

# Constructing and Mapping Fuzzy Thematic Clusters to Higher Ranks in a Taxonomy

Boris Mirkin<sup>1, 2</sup>, Susana Nascimento<sup>3</sup>, Trevor Fenner<sup>1</sup>, and Luís Moniz Pereira<sup>3</sup>

<sup>1</sup> School of Computer Science, Birkbeck University of London, London, WC1E 7HX, UK,

<sup>2</sup> Division of Applied Mathematics, Higher School of Economics, Moscow, RF

<sup>3</sup> Computer Science Department and Centre for Artificial Intelligence (CENTRIA), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

**Abstract.** We present a method for mapping a structure such as a research department to a related taxonomy in a thematically consistent way. The components of the structure are supplied with fuzzy membership profiles over the taxonomy. The profiles are then generalized in two steps: first, by fuzzy clustering, and then by mapping the clusters to higher ranks of the taxonomy. To be specific, we concentrate on the Computer Sciences area represented by the taxonomy of ACM Computing Classification System (ACM-CCS). We build fuzzy clusters of the taxonomy leaves according to the similarity between individual profiles. Clusters are extracted using an original additive spectral clustering method involving a number of model-based stopping conditions. The clusters are not necessarily consistent with the taxonomy. This is formalized by parsimoniously lifting them to higher ranks of the taxonomy using an original recursive algorithm for minimizing a penalty function that involves “head subjects” on the higher ranks of the taxonomy along with their “gaps” and “offshoots”. An example is given illustrating the method applied to real-world data.

## 1 Introduction

The last decade has witnessed an unprecedented rise of the concept of ontology as a computationally feasible tool for knowledge maintenance. For example, the usage of Gene Ontology [6] for interpretation and annotation of various gene sets and gene expression data is becoming a matter of routine in bioinformatics (see, for example, [14] and references therein).

To apply similar approaches to less digitalized domains, such as activities of organizations in a field of science or knowledge supplied by teaching courses in a university school, one needs to distinguish between different levels of data and knowledge, and build techniques for deriving and transforming corresponding bits of knowledge within a comprehensible framework.

The goal of this paper is to present such a framework built on top of a pre-specified taxonomy of the domain under consideration as the base. In general, a taxonomy is a rooted-tree-like structure whose nodes correspond to individual topics in such a way that the parental node’s topic generalizes the topics of its children’s nodes. We concentrate on the issue of representing an organization or any other system under consideration, in terms of the taxonomy topics. We first build profiles for its constituent entities in terms of the taxonomy and then thematically generalize them.

To represent a functioning structure over a taxonomy is to indicate those topics in the taxonomy that most fully express the structure's working in its relation to the taxonomy. It may seem that conventionally mapping the system to all nodes related to topics involved in the profiles within the structure would do the job, but it is not the case - such a mapping typically represents a fragmentary set of many nodes without any clear picture of thematic interrelation among them. Therefore, to make the representation thematically consistent and parsimonious, we propose a two-phase generalization approach. The first phase generalizes over the structure by building clusters of taxonomy topics according to the functioning of the system. The second phase generalizes over the clusters by parsimoniously mapping them to higher ranks of the taxonomy according to a special parsimonious "lift" procedure. Both entity profiles and thematic clusters derived at the first phase are fuzzy in order to better reflect the real world objects, so that the lifting method applies to fuzzy clusters. It should be pointed out that both building fuzzy profiles and finding fuzzy clusters are research activities well documented in the literature; yet the issues involved in this project led us to develop some original schemes of our own including an efficient method for fuzzy clustering combining the approaches of spectral and approximation clustering [12].

We apply these constructions in two areas: (i) to visualize activities of Computer Science research organizations; and (ii) to discern the complexes of mathematical ideas according to classes taught in regular teaching courses in a university department. We take the popular ACM Computing Classification System (ACM-CCS), a conceptual four-level classification of the Computer Science subject area as a pre-specified taxonomy for (i), and the three-layer Mathematics Subject Classification MSC2010 developed by the Mathematical Reviews and Zentralblatt Mathematics editors (see <http://www.ams.org/mathscinet/msc/msc2010.html>), for (ii). In what follows the focus is mainly on the application (i) to research organizations. The paper is organized according to the structure of our approach: Section 2 describes an e-system we developed for getting ACM-CCS leaves fuzzy membership profiles from Computer Science researchers, Section 3 describes our method for deriving fuzzy clusters from the profiles, and Section 4 presents our parsimonious lift method to generalize to higher ranks in a taxonomy tree.

## 2 Taxonomy-based profiles

### 2.1 Representing over the ACM-CCS taxonomy

In the case of investigation of activities of a university department or center, a research team's profile can be defined as a fuzzy membership function on the set of leaf-nodes of the taxonomy under consideration so that the memberships reflect the extent of the team's effort put into corresponding research topics.

In this case, the ACM Computing Classification System (ACM-CCS) [1] is used as the taxonomy. ACM-CCS comprises eleven major partitions (first-level subjects) such as *B. Hardware*, *D. Software*, *E. Data*, *G. Mathematics of Computing*, *H. Information Systems*, etc. These are subdivided into 81 second-level subjects. For example, item *I. Computing Methodologies* consists of eight subjects including *I.1 SYMBOLIC AND ALGEBRAIC MANIPULATION*, *I.2 ARTIFICIAL INTELLIGENCE*, *I.5 PATTERN*

*RECOGNITION*, etc. They are further subdivided into third-layer topics as, for instance, *I.5 PATTERN RECOGNITION* which is represented by seven topics including *I.5.3 Clustering*, *I.5.4 Applications*, etc.

Taxonomy structures such as the ACM-CCS are used, mainly, as devices for annotation and search for documents or publications in collections such as that on the ACM portal [1]. The ACM-CCS tree has been applied also as: a gold standard for ontologies derived by web mining systems such as the CORDER engine [17]; a device for determining the semantic similarity in information retrieval [9] and e-learning applications [18, 5]; and a device for matching software practitioners' needs and software researchers' activities [4].

Here we concentrate on a different application of ACM-CCS – a generalized representation of a Computer Science research organization that can be used for overviewing scientific subjects that are being developed in the organization, assessing the scientific issues in which the character of activities in organizations does not fit well onto the classification – these can potentially be the growth points, and help with planning the restructuring of research and investment.

## 2.2 E-Screen survey tool

Fuzzy profiles are derived from either automatic analysis of documents posted on the web by the teams or by explicitly surveying the members of the department. The latter option is especially convenient in situations in which the web contents do not properly reflect the developments, for example, in non-English speaking countries with relatively underdeveloped internet infrastructures for the maintenance of research results. We developed an interactive survey tool that provides two types of functionality: i) collection of data about ACM-CCS based research profiles of individual members; ii) statistical analysis and visualization of the data and results of the survey on the level of a department. The respondent is asked to select up to six topics among the leaf nodes of the ACM-CCS tree and assign each with a percentage expressing the proportion of the topic in the total of the respondent's research activity for, say, the past four years. Figure 1 shows a screenshot of the baseline interface for a respondent who has chosen six ACM-CCS topics during her survey session. A deeper view can be taken with a different "research results" form that allows to make a more detailed assessment in terms of individual research results in several categories such as refereed publications, funded projects, and theses supervised. These can be used for weighting research results with respect to the topics in ACM-CCS.

The set of profiles supplied by respondents forms an  $N \times M$  matrix  $F$  where  $N$  is the number of ACM-CCS topics involved in the profiles and  $M$  the number of respondents. Each column of  $F$  is a fuzzy membership function, rather sharply delineated because only six topics may have positive memberships in each of the columns.

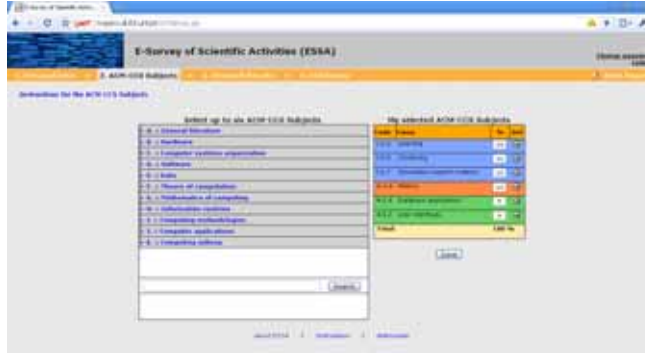


Fig. 1: Screenshot of the interface survey tool for selection of ACM-CCS topics.

### 3 Representing research organization by fuzzy clusters of ACM-CCS topics

#### 3.1 Deriving similarity between ACM-CCS research topics

We represent a research organization by clusters of ACM-CCS topics to reflect thematic communalities between activities of members or teams working on these topics. The clusters are found by analyzing similarities between topics according to their appearances in the profiles. The more profiles contain a pair of topics  $i$  and  $j$  and the greater the memberships of these topics, the greater is the similarity score for the pair.

In spite of the fact that many fuzzy clustering algorithms have been developed already [2], [7], most of them are ad hoc and, moreover, it is not straightforward to apply them in real because they all involve manually specified parameters such as the number of clusters or threshold of similarity. We apply a model-based approach of additive clustering, which is combined with the spectral clustering approach to make it practical and derive model-based parameters helping to choose the right number of clusters.

Consider a set of  $V$  individuals ( $v = 1, 2, \dots, V$ ), engaged in research over some topics  $t \in T$  where  $T$  is a pre-specified set of scientific subjects. The level of research effort by individual  $v$  in developing topic  $t$  is evaluated by the membership  $f_{tv}$  in profile  $f_v$  ( $v = 1, 2, \dots, V$ ).

Then the similarity  $w_{tt'}$  between topics  $t$  and  $t'$  is defined as

$$w_{tt'} = \sum_{v=1}^V \frac{n_v}{n_{max}} f_{tv} f_{t'v}, \quad (1)$$

where the ratios are introduced to balance the scores of individuals bearing different numbers of topics.

To make the cluster structure in the similarity matrix sharper, we apply the spectral clustering approach to pre-process the similarity matrix  $W$  using the so-called Laplacian transformation [8]. First, an  $N \times N$  diagonal matrix  $D$  is defined, with  $(t, t)$  entry

equal to  $d_t = \sum_{t' \in T} w_{tt'}$ , the sum of  $t$ 's row of  $W$ . Then unnormalized Laplacian and normalized Laplacian are defined by equations  $L = D - W$  and  $L_n = D^{-1/2} L D^{-1/2}$ , respectively. Both matrices are semipositive definite and have zero as the minimum eigenvalue. The minimum non-zero eigenvalues and corresponding eigenvectors of the Laplacian matrices are utilized then as relaxations of combinatorial partition problems [16, 8]. Of comparative properties of these two normalizations, the normalized Laplacian, in general, is considered superior [8]. Since the additive clustering approach described in the next section relies on maximum rather than minimum eigenvalues, we use the Laplacian pseudoinverse transformation, Lapin for short, defined by

$$L_n^-(W) = \tilde{Z} \tilde{\Lambda}^{-1} \tilde{Z}'$$

where  $\tilde{\Lambda}$  and  $\tilde{Z}$  are defined by the spectral decomposition  $L_n = Z \Lambda Z'$  of matrix  $L_n = D^{-1/2}(D - W)D^{-1/2}$ . To specify these matrices, first, set  $T'$  of indices of elements corresponding to non-zero elements of  $\Lambda$  is determined, after which the matrices are taken as  $\tilde{\Lambda} = \Lambda(T', T')$  and  $\tilde{Z} = Z(:, T')$ . The choice of the Lapin transformation can be explained by the fact that it leaves the eigenvectors of  $L_n$  unchanged while inverting the non-zero eigenvalues  $\lambda \neq 0$  to those  $1/\lambda$  of  $L_n^-$ . Then the maximum eigenvalue of  $L_n^-$  is the inverse of the minimum non-zero eigenvalue  $\lambda_1$  of  $L_n$ , corresponding to the same eigenvector.

### 3.2 Additive fuzzy clusters using a spectral method

Thematic similarities  $a_{tt'}$  between topics are but manifested expressions of some hidden patterns within the organization which can be represented by fuzzy clusters in exactly the same manner as the manifested scores in the definition of the similarity  $w_{tt'}$  (1). We assume that a thematic fuzzy cluster is represented by a membership vector  $u = (u_t)$ ,  $t \in T$ , such that  $0 \leq u_t \leq 1$  for all  $t \in T$ , and an intensity  $\mu > 0$  that expresses the extent of significance of the pattern corresponding to the cluster, within the organization under consideration. With the introduction of the intensity, applied as a scaling factor to  $u$ , it is the product  $\mu u$  that is a solution rather than its individual co-factors. Given a value of the product  $\mu u_t$ , it is impossible to tell which part of it is  $\mu$  and which  $u_t$ . To resolve this, we follow a conventional scheme: let us constrain the scale of the membership vector  $u$  on a constant level, for example, by a condition such as  $\sum_t u_t = 1$  or  $\sum_t u_t^2 = 1$ , then the remaining factor will define the value of  $\mu$ .

Our additive fuzzy clustering model follows that of [15, 10, 13] and involves  $K$  fuzzy clusters that reproduce the pseudo-inverted Laplacian similarities  $a_{tt'}$  up to additive errors according to the following equations:

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

where  $u_k = (u_{kt})$  is the membership vector of cluster  $k$ , and  $\mu_k$  its intensity.

The item  $\mu_k^2 u_{kt} u_{kt'}$  expresses the contribution of cluster  $k$  to the similarity  $a_{tt'}$  between topics  $t$  and  $t'$ , which depends on both the cluster's intensity and the membership values. The value  $\mu^2$  summarizes the contribution of intensity and will be referred to as the cluster's weight.

To fit the model in (2), we apply the least-squares approach, thus minimizing the sum of all  $e_{tt'}^2$ . Since  $A$  is definite semi-positive, its first  $K$  eigenvalues and corresponding eigenvectors form a solution to this if no constraints on vectors  $u_k$  are imposed. Additionally, we apply the one-by-one principal component analysis strategy for finding one cluster at a time this makes the computation feasible and is crucial for determining the number of clusters. Specifically, at each step, we consider the problem of minimization of a reduced to one fuzzy cluster least-squares criterion

$$E = \sum_{t,t' \in T} (b_{tt'} - \xi u_t u_{t'})^2 \quad (3)$$

with respect to unknown positive  $\xi$  weight (so that the intensity  $\mu$  is the square root of  $\xi$ ) and fuzzy membership vector  $u = (u_t)$ , given similarity matrix  $B = (b_{tt'})$ .

At the first step,  $B$  is taken to be equal to  $A$ . Each found cluster changes  $B$  by subtracting the contribution of the found cluster (which is additive according to model (2)), so that the residual similarity matrix for obtaining the next cluster will be  $B - \mu^2 uu^T$  where  $\mu$  and  $u$  are the intensity and membership vector of the found cluster. In this way,  $A$  indeed is additively decomposed according to formula (2) and the number of clusters  $K$  can be determined in the process.

Let us specify an arbitrary membership vector  $u$  and find the value of  $\xi$  minimizing criterion (3) at this  $u$  by using the first-order condition of optimality:

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

so that the optimal  $\xi$  is

$$\xi = \frac{\mathbf{u}' B \mathbf{u}}{(\mathbf{u}' \mathbf{u})^2} \quad (4)$$

which is obviously non-negative if  $B$  is semi-positive definite.

By putting this  $\xi$  in equation (3), we arrive at

$$E = \sum_{t,t' \in T} b_{tt'}^2 - \xi^2 \sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2 = S(B) - \xi^2 (\mathbf{u}' \mathbf{u})^2,$$

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

Let us denote the last item by

$$G(u) = \xi^2 (\mathbf{u}' \mathbf{u})^2 = \left( \frac{\mathbf{u}' B \mathbf{u}}{\mathbf{u}' \mathbf{u}} \right)^2, \quad (5)$$

so that the similarity data scatter is the sum:

$$S(B) = G(u) + E \quad (6)$$

of two parts,  $G(u)$ , which is explained by cluster  $(\mu, u)$ , and  $E$ , which remains unexplained.

An optimal cluster, according to (6), is to maximize the explained part  $G(u)$  in (5) or its square root

$$g(u) = \xi \mathbf{u}'\mathbf{u} = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'\mathbf{u}}, \quad (7)$$

which is the celebrated Rayleigh-Ritz quotient, whose maximum value is the maximum eigenvalue of matrix  $B$ , which is reached at its corresponding eigenvector, in the unconstrained problem.

This shows that the spectral clustering approach is appropriate for our problem. According to this approach, one should find the maximum eigenvalue  $\lambda$  and corresponding normed eigenvector  $z$  for  $B$ ,  $[\lambda, z] = A(B)$ , and take its projection to the set of admissible fuzzy membership vectors.

According to this approach, a number of model-based criteria emerges for halting the process of sequential extraction of fuzzy clusters:

1. The optimal value of  $\xi$  (4) for the spectral fuzzy cluster becomes negative.
2. The contribution of a single extracted cluster becomes too low, less than a pre-specified  $\tau > 0$  value.
3. The residual scatter  $E$  becomes smaller than a pre-specified  $\epsilon$  value, say less than 5% of the original similarity data scatter.

The described one-by-one fuzzy thematic cluster extraction spectral algorithm is referred to as the ADDI-FS. Its effectiveness has been demonstrated by either theoretical or experimental considerations or both, versus each of the three categories of clustering approaches - additive clustering, spectral clustering, and relational fuzzy clustering - it relates to. In the context of additive clustering, fuzzy approaches were considered only by [13] in a very restricted setting: (a) the clusters intensities are assumed constant there, (b) the number of clusters is pre-specified, and (c) the fitting method is very local and computationally intensive - these restrictions are overcome in ADDI-FS. The genuine spectral clustering [16, 8] applies only to crisp clustering, yet some modified formats are available such as [19]; ADDI-FS applied to problems of the analysis of community structure such as "Zachary karate club" outperformed the method proposed in [19]. Regarding the relational fuzzy clustering approaches [2], we demonstrated the effectiveness of ADDI-FS by comparing its results on data of Gaussian clusters generated as described in [3] with those obtained in [3] in the experiments with five relational fuzzy clustering algorithms: measured by the adjusted Rand index, the ADDI-FS cluster recovery was 0.71 against 0.67 as the best result reported in [3] (see details in [12]).

## 4 Parsimonious Lifting Method

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 (see Fig. 2). Depending on the extent of inconsistency between the cluster and the taxonomy, such lifting can be done differently, leading to different portrayals of the cluster on ACM-CCS tree depending on the relative weights of the

events taken into account. A major event is the so-called “head subject”, a taxonomy node covering (some of) leaves belonging to the cluster, so that the cluster is represented by a set of head subjects. The penalty of the representation to be minimized is proportional to the number of head subjects so that the smaller that number the better. Yet the head subjects cannot be lifted too high in the tree because of the penalties for associated events, the cluster “gaps” and “offshoots” the number of them depends on the extent of inconsistency of the cluster versus the taxonomy.

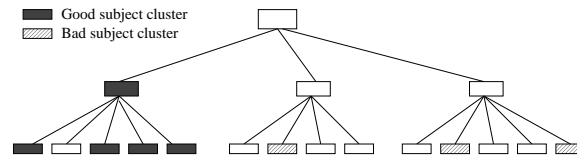


Fig. 2: Two clusters of second-layer topics, presented with checked and diagonal-lined boxes, respectively. The checked box cluster fits within one first-level category (with one gap only), whereas the diagonal line box cluster is dispersed among two categories on the right. The former fits the classification well; the latter does not.

The gaps are head subject’s children topics that are not included in the cluster. An offshoot is a taxonomy leaf node that is a head subject (not lifted). It is not difficult to see that the gaps and offshoots are determined by the head subjects specified in a lift (see Fig. 3).

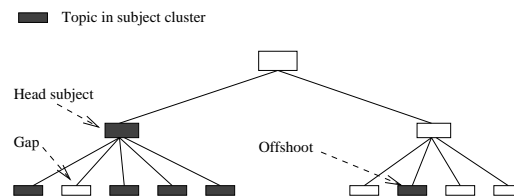


Fig. 3: Three types of features in mapping of a subject cluster to the taxonomy.

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, as illustrated on Fig. 5 later.

Altogether, the set of topic clusters together with their optimal head subjects, offshoots and gaps constitute a parsimonious representation of the organization. Such a representation can be easily accessed and expressed. It can be further elaborated by

highlighting those subjects in which members of the organization have been especially successful (i.e., publication in best journals or awards) or distinguished by a special feature (i.e., industrial use or inclusion in a teaching program). Multiple head subjects and offshoots, when they persist at subject clusters in different organizations, may show some tendencies in the development of the science, that the classification has not taken into account yet.

A parsimonious lift of a subject cluster can be achieved by recursively building a parsimonious representation for each node of the ACM-CCS tree based on parsimonious representations for its children. In this, we assume that any head subject is automatically present at each of the nodes it covers, unless they are gaps (as presented on Fig. 3). Our algorithm is set as a recursive procedure over the tree starting at leaf nodes.

The procedure determines, at each node of the tree, sets of head gain and gap events to iteratively raise them to those of the parents, under each of two different assumptions that specify the situation at the parental node. One assumption is that the head subject has been inherited at the parental node from its own parent, and the second assumption is that it has not been inherited but gained in the node only. In the latter case the parental node is labeled as a head subject. Consider the parent-children system as shown in Fig. 4, with each node assigned with sets of gap and head gain events under the above two inheritance of head subject assumptions.

Let us denote the total penalty, to be minimized, under the inheritance and non-inheritance assumptions by  $p_i$  and  $p_n$ , respectively. A lifting result at a given node is defined by a pair of sets (H, G), representing the tree nodes at which events of head gains and gaps, respectively, have occurred in the subtree rooted at the node. We use  $(H_i, G_i)$  and  $(H_n, G_n)$  to denote lifting results under the inheritance and non-inheritance assumptions, respectively. The algorithm computes parsimonious representations for parental nodes according to the topology of the tree, proceeding from the leaves to the root in the manner which is similar to that described in [11] for a mathematical problem in bioinformatics.

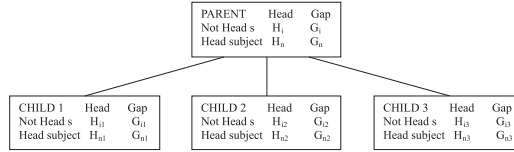


Fig. 4: Events in a parent-children system according to a parsimonious lift scenario.

For the sake of simplicity, we present only a version of the algorithm for crisp clusters obtained by a defuzzification step. Given a crisp topic cluster  $S$ , and penalties  $h$ ,  $o$  and  $g$  for being a head subject, offshoot and gap, respectively, the algorithm is initialized as follows.

At each leaf  $l$  of the tree, either  $H_n = \{l\}$ , if  $l \in S$ , or  $G_i = \{l\}$ , otherwise. The other three sets are empty. The penalties associated are  $p_i = 0$ ,  $p_n = o$  if  $H_n$  is not

empty, that is, if  $l \in S$ , and  $p_i = g$ ,  $p_n = 0$ , otherwise. This is obviously a parsimonious arrangement at the leaf level.

The recursive step applies to any node  $t$  whose children  $v \in V$  have been assigned with the two couples of  $H$  and  $G$  sets already (see Figure 4 at which  $V$  consists of three children):  $(H_i(v), L_i(v); H_n(v), L_n(v))$  along with associated penalties  $p_i(v)$  and  $p_n(v)$ .

(I) Deriving the pair  $H_i(t)$  and  $G_i(t)$ , under the inheritance assumption, the one of the following two cases is to be chosen depending on the cost:

(a) The head subject has been lost at  $t$ , so that  $H_i(t) = \cup_{v \in V} H_n(v)$  and  $G_i(t) = \cup_{v \in V} G_n(v) \cup \{t\}$ . (Note different indexes,  $i$  and  $n$  in the latter expression.) The penalty in this case is  $p_i = \sum_{v \in V} p_n(v) + g$ ;

or

(b) The head subject has not been lost at  $t$ , so that  $H_i(t) = \emptyset$  (under the assumption that no gain can happen after a loss) and  $G_i = \cup_{v \in V} G_i(v)$  with  $p_i = \sum_{v \in V} p_i(v)$ .

The case that corresponds to the minimum of the two  $p_i$  values is returned then.

(II) Deriving the pair  $H_n(t)$  and  $G_n(t)$ , under the non-inheritance assumption, the one of the following two cases is to be chosen that minimizes the penalty  $p_n$ :

(a) The head subject has been gained at  $t$ , so that  $H_n(t) = \cup_{v \in V} H_i(v) \cup \{t\}$  and  $G_n(t) = \cup_{v \in V} G_i(v)$  with  $p_n = \sum_{v \in V} p_i(v) + h$ ;

or (b) The head subject has not been gained at  $t$ , so that  $H_n(t) = \cup_{v \in V} H_n(v)$  and  $G_n = \cup_{v \in V} G_n(v)$  with  $p_n = \sum_{v \in V} p_n(v)$ .

After all tree nodes  $t$  have been assigned with the two pairs of sets, accept the  $H_n$ ,  $L_n$  and  $p_n$  at the root. This gives a full account of the events in the tree.

This algorithm leads indeed to an optimal representation; its extension to a fuzzy cluster is achieved through using the cluster memberships in computing the penalty values.

## 5 An application to a real world case

Let us illustrate the approach by using the data from a survey conducted at the Centre of Artificial Intelligence, Faculty of Science & Technology, New University of Lisboa (CENTRIA-UNL). The survey involved 16 members of the academic staff of the Centre who covered 46 topics of the third layer of the ACM-CCS.

With the algorithm ADDI-FS applied to the  $46 \times 46$  similarity matrix, two clusters have been sequentially extracted, after which the residual matrix has become definite negative (stopping condition (a)). Cluster 1 is of pattern recognition and its applications to physical sciences and engineering including images and languages, with offshoots to general aspects of information systems. In cluster 2, all major aspects of computational mathematics are covered, with an emphasis on reliability and testing, and with applications in the areas of life sciences. Overall these results are consistent with the informal assessment of the research conducted in the research organization. Moreover, the sets of research topics chosen by individual members at the ESSA survey follow the cluster structure rather closely, falling mostly within one of the two.

Figure 5 shows the representation of CENTRIA's cluster 2 in the ACM-CCS taxonomy with penalties of  $h = 1$ ,  $o = 0.8$ , and  $g = 0.1$ .

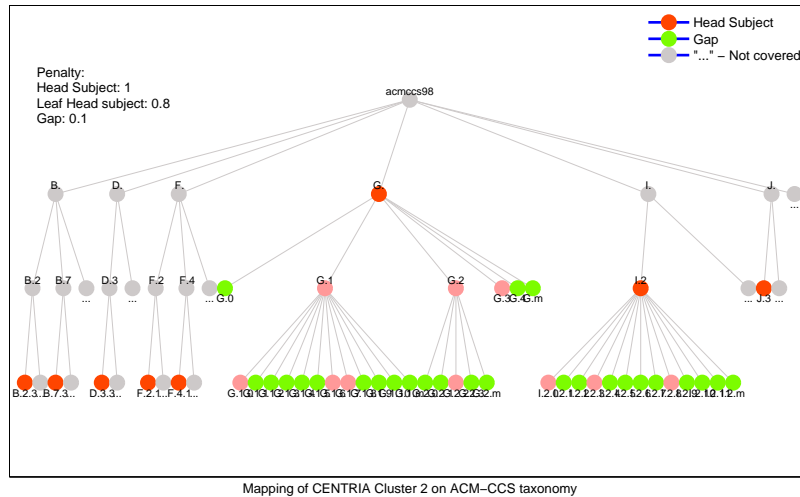


Fig. 5: Mapping of CENTRIA cluster 2 onto the ACM-CCS tree with penalties  $h = 1$ ,  $o = 0.8$  and  $g = 0.1$ : two head subjects along with 10 and 16 gaps, respectively.

## 6 Conclusion

We have described a method for knowledge generalization that employs a taxonomy tree. The method constructs fuzzy membership profiles of the entities constituting the structure under consideration in terms of the taxonomies leaves, and then it generalizes them in two steps. These steps are:

- (i) fuzzy clustering research topics according to their thematic similarities, ignoring the topology of the taxonomy, and
- (ii) lifting clusters mapped to the taxonomy to higher ranked categories in the tree.

These generalization steps thus cover both sides of the representation process: the empirical – related to the structure under consideration – and the conceptual – related to the taxonomy hierarchy.

Potentially, this approach could lead to a useful instrument for comprehensive visual representation of developments in any field of organized human activities.

## Acknowledgments

The authors are grateful to CENTRIA-UNL members that participated in the survey. Igor Guerreiro is acknowledged for developing software for the ESSA tool. Rui Felizardo is acknowledged for developing software for the lifting algorithm with interface shown in Figures 5. This work has been supported by grant PTDC/EIA/69988/2006 from the Portuguese Foundation for Science & Technology. The support of the individual research project 09-01-0071 “Analysis of relations between spectral and approximation clustering” to BM by the “Science Foundation” Programme of the State University – Higher School of Economics, Moscow RF, is also acknowledged.

## References

1. *ACM Computing Classification System*, 1998, <http://www.acm.org/about/class/1998>. Cited 9 Sep 2008.
2. Bezdek, J., Keller, J., Krishnapuram, R., Pal, T.: *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, (1999)
3. Brouwer, R.: A method of relational fuzzy clustering based on producing feature vectors using FastMap, *Information Sciences*, 179, pp. 3561-3582 (2009)
4. Feather, M., Menzies, T., Connelly, J.: "Matching software practitioner needs to researcher activities", *Proc. of the 10th Asia-Pacific Software Engineering Conference (APSEC'03)*, IEEE, pp. 6, (2003)
5. Gaevic, D., Hatala, .: "Ontology mappings to improve learning resource search", *British Journal of Educational Technology*, 37(3), pp. 375 - 389, (2006)
6. "The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology", *Nature Genetics*, 25, pp. 25-29 (2000)
7. Liu, J., Wang, W., Yang, J.: "Gene ontology friendly biclustering of expression profiles", *Proc. of the IEEE Computational Systems Bioinformatics Conference*. IEEE, pp. 436-447, (2004)
8. von Luxburg, U.: A tutorial on spectral clustering, *Statistics and Computing* 17, pp. 395-416 (2007)
9. Miralaei, S., Ghorbani, A.: "Category-based similarity algorithm for semantic similarity in multi-agent information sharing systems", *IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology*, pp. 242-245 (2005)
10. Mirkin, B.: "Additive clustering and qualitative factor analysis methods for similarity matrices", *Journal of Classification*, 4(1), pp. 7-31, (1987)
11. Mirkin, B., Fenner, T., Galperin, M., Koonin, E.: "Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes", *BMC Evolutionary Biology*, 3:2, (2003)
12. Mirkin, B., Nascimento, S.: "Analysis of Community Structure, Affinity Data and Research Activities using Additive Fuzzy Spectral Clustering", Technical Report 6, School of Computer Science, Birkbeck University of London (2009)
13. Sato, M., Sato, Y., Jain, L.C.: *Fuzzy Clustering Models and Applications*, Physica-Verlag, Heidelberg, (1997)
14. Skarman, A., Jiang, L., Hornshoj, H., Buitenhuis, B., Hedegaard, J., Conley, L., Sorensen, P.: "Gene set analysis methods applied to chicken microarray expression data", *BMC Proceedings*, 3(Suppl 4) (2009)
15. Shepard, R.N., Arabie, P.: "Additive clustering: representation of similarities as combinations of overlapping properties", *Psychological Review* 86, 87-123, (1979)
16. Shi, J., Malik, J.: "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888-905 (2000)
17. Thorne, C., Zhu, J., Uren, V.: "Extracting domain ontologies with CORDER", *Tech. Reportkmi-05-14*. Open University, 1-15 (2005)
18. Yang, L., Ball, M., Bhavsar, V., Boley, H.: "Weighted partonomy-taxonomy trees with local similarity measures for semantic buyer-seller match-making", *Journal of Business and Technology*. Atlantic Academic Press, 1(1), 42-52 (2005)
19. Zhang, S., Wang, R.-S., Zhang, X.-S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering, *Physica A* 374, 483-490 (2007)