Representing a Computer Science Research Organization on the ACM Computing Classification System

Boris Mirkin[1], Susana Nascimento[2], and Luis Moniz Pereira[2]

[1] School of Computer Science
Birkbeck University of London
London, UK WC1E 7HX
[2] Computer Science Department and Centre for Artificial Intelligence (CENTRIA)
FCT, Universidade Nova de Lisboa
Caparica, Portugal

Abstract. We propose a method, Cluster-Lift, for parsimoniously mapping clusters of ontology classes of lower levels onto a subset of high level classes in such a way that the latter can be considered as a generalized description of the former. Specifically, we consider the problem of visualization of activities of a Computer Science Research organization on the ACM Computing Subjects Classification (ACMC), which is a three level taxonomy.
It is possible to specify the set of ACMC subjects that are investigated by the organization's teams and individual members and map them to the ACMC hierarchy. This visualization, however, usually appears overly detailed, confusing, and difficult to interpret. This is why we propose a two-stage Cluster-Lift procedure. On the first stage, the subjects are clustered according to their similarity defined in such a way that the greater the number of researchers working on a pair of subjects, the greater the similarity between the pair. On the second stage, each subject cluster is mapped onto ACMC and lifted within the taxonomy. The lifting involves a formalization of the concept of "head subject", as well as its "gaps" and "offshoots" and is to be done in a parsimonious way by minimizing a weighted sum of the numbers of head subjects, gaps and offshoots. The Cluster-Lift results are easy to see and interpret.
A real-world example of the working of our approach is provided.

1 ACM Computing Classification System Fits for Representing CS Research Activities

ACM Computing Classification System (ACMC) is a conceptual three level classification of the Computer Science subject area built to reflect the vast and changing world of computer oriented writing. This classification was first published in 1982 and then thoroughly revised in 1998 and it is being revised since [1]. The ACMC is used, mainly, as a device for annotation and search for publications in collections such as that on the ACM portal [1], that is, for the library and bibliographic applications. Here we propose its use for representing research organizations in such a way that the organization's research topics are generalized by parsimoniously lifting them, after clustering, along the ACMC topology.

Potentially, this kind of ACMC representation can be used for the following purposes:

i Overview of scientific subjects being developed in an organization.
ii Positioning the organization over ACMC.
iii Overview of scientific disciplines being developed in organizations over a country or other territorial unit, with a quantitative assessment of controversial subjects, for example, those in which the level of activity is not sufficient or the level of activities by far excesses the level of results.
iv Assessing the scientific issues in which the character of activities in organizations does not fit well onto the classification; these can be potentially the growth points or other breakthrough developments.
v Planning research restructuring and investment.

2 Cluster - Lift Method

We represent a research organization by clusters of ACMC topics emerging according to members or teams simultaneously working on them. Each of the clusters is mapped to the ACMC tree and then lifted in the tree to express its general tendencies. The clusters are found by analyzing similarities between topics as derived from either automatic analysis of documents posted on web by the teams or by explicitly surveying the members of the department. The latter option is especially convenient at situations in which the web contents do not properly reflect the developments. Then we need a survey tool.

Accordingly, this work involves developing the following tools. 1) A e-screen based ACMC topic surveying device. 2) A method for deriving similarity between ACMC topics. 3) A robust method for finding possibly overlapping subject clusters. 4) A method for parsimoniously lifting topic clusters on ACMC. In the following subsections, we describe them in turn.

2.1 E-screen survey tool

An interactive survey tool has been developed to provide two types of functionalities about the research activities in an organization: i) data collection about the research results of individual members, described according to the ACMC topics; ii) statistical analysis and visualization of the data and results of the survey. The period of research activities comprises the survey year and the previous four years. This is supplied with simultaneous “focus + context” navigation functionalities as well as quick interaction with the taxonomy [2]. The respondent is asked to select up to six topics in the third layer of the ACMC tree and assign each with a percentage expressing the proportion of the topic in the total of the respondent’s research activity. Figure 1 shows a screenshot of the interface for a respondent who chose six ACMC topics during his/her survey session. Another, “research results” form allows to make a more detailed assessment in terms of research results of the respondent in categories such as refereed publications, funded projects, and theses supervised.

The (third-layer) nodes of the ACMC tree are populated thus by respondents’ weights, which can be interpreted as membership degrees of the respondent’s activity to the ACMC topics.

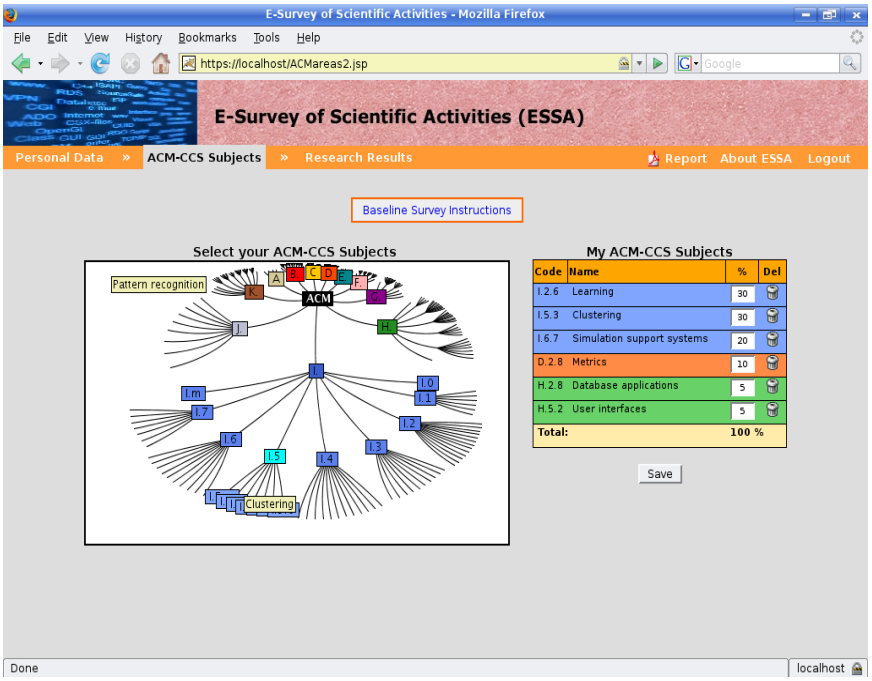

Fig. 1. Screenshot of the interface survey tool to select ACMC topics.

2.2 Deriving similarity between ACMC topics

We derive similarity between ACMC topics i and j as the weighted sum of individual similarities. The individual similarity is just the product of weights f_i and f_j assigned by the respondent to the topics. Clearly, topics that are left outside of the individual's list, have zero similarities with other topics.

The individual's weight is inversely proportional to the number of subjects they selected in the survey. This smoothes out the differences between topic weights imposed by the selection sizes.

It is not difficult to see that the resulting topic-to-topic similarity matrix $A = (a_{ij})$ is positive semidefinite.

2.3 Finding overlapping clusters

The issue of determining of the subject clusters can be explicated as the well-known problem of finding clusters, potentially overlapping, over similarity matrix $A = (a_{ij})$.

We employ for this the data recovery approach described in [3] for the case of crisp clustering and in [4] for the case of fuzzy clustering. Here we consider only the crisp clustering case.

Let us denote $s = (s_i)$ a binary membership vector defining a subset of ACMC topics $S = \{i : s_i = 1\}$. The one-cluster criterion to be optimized by the cluster S to be found is expressed as:

$$g(S) = s^T As / s^T s = a(S)|S|. \quad (1)$$

where $a(S)$ is the average similarity a_{ij} within S and $|S|$ the number of entities in S. This criterion has a simple intuitive meaning as a compromise between too contradicting criteria: (a) maximizing the within-cluster similarity and (b) maximizing the cluster size. When squared, the criterion expresses the proportion of the data scatter, which is taken into account by cluster S according to the data recovery model described in [3].

It should be pointed out that this criterion not only emerges in the data recovery framework but it also fits into some other frameworks such as (i) maximum density subgraphs [6] and (ii) spectral clustering [7].

ADDI-S algorithm starts from $S = \{i\}$ where i is any topic $i \in I$, and, in this way, produces a number of potentially overlapping or even coinciding locally optimal clusters S_i – these are considered then for selection according to their contribution weights $g(S_i)^2$ and the extent of their overlap with the other clusters. The intuition behind this heuristic is that each of the locally optimal clusters is well separated from the rest; therefore, a small number of them covering a major part of the data set is a good representation of the similarities.

The algorithm iteratively finds an entity $j \notin S$ by maximizing $g(S \pm j)$ where $S \pm j$ stands for $S + j$ if $j \notin S$ or $S - j$ if $j \in S$. It appears, for doing this one just needs to compare the average similarity between j and S with the threshold $\pi = a(S)/2$. Obviously, the produced S is rather tight because each $i \in S$ has a high degree of similarity with S, greater than half of the average similarity within S, and simultaneously is well separated from the rest, because for each entity $j \notin S$, its average similarity with S is less than that.

2.4 Parsimonious lifting method

To generalise the main contents of a subject cluster, we translate it to higher layers of the taxonomy by lifting it according to the principle: if all or almost all children of a node belong to the cluster, then the node represents the cluster on a higher level of the ACMC taxonomy. Such a lift can be done differently leading to different portrayals of that on the ACMC tree. A cluster can fit quite well into the classification or not (see Figure 2), depending on how much its topics are dispersed among the tree nodes.

The best possible fit would be when all topics in the subject cluster fall within a parental node in such a way that all the siblings are covered and no gap occurs. The parental tree node, in this case, can be considered as the head subject of the cluster. A few gaps, that is, head subject's children topics that are not included in the cluster, although diminish the fit, still leave the head subject unchanged. A larger misfit occurs when a cluster is dispersed among two or more head subjects. One more type of misfit may emerge when almost all cluster topics fall within the same head subject node but one or two of the topics offshoot to other parts of the classification tree (see Figure 3).

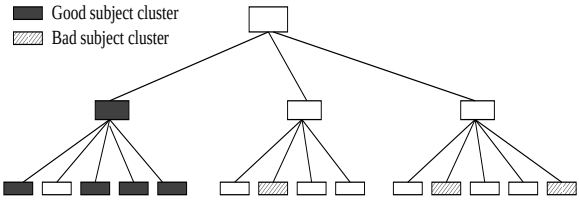

Fig. 2. Two clusters of second-layer topics, presented with checked and diagonal lined boxes, respectively. The check box cluster fits all within one first -level category (with one gap only), whereas the diagonal line box cluster is dispersed among two categories on the right. The former fits the classification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

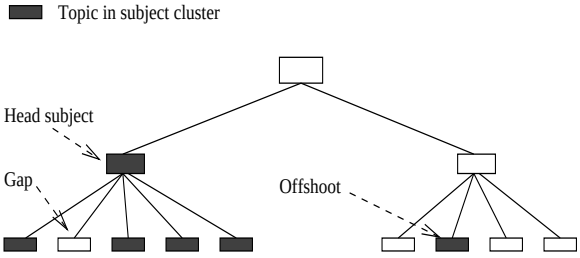

Fig. 3. Three types of features of mapping of a subject cluster to the ontology.

Such offshoots, when persist at subject clusters in different organizations, may show some tendencies in the development of the science, that the classification tree has not taken into account yet. The total count of head subjects, gaps and offshoots, each type weighted accordingly, can be used for scoring the extent of effort needed for lifting a research grouping over classification tree as illustrated on Figure 4. The smaller the score, the better the fit. When the topics under consideration relate to deeper levels of classification, such as the third layer of ACMC, the scoring may allow some tradeoff between different possibilities for lifting clusters to the head subjects. As illustrated on Figure 4, the subject cluster of third-layer topics presented by checked boxes, can be lifted to two head subjects as on (A) or, just one, the upper category on (B), with the “cost” of three more gap nodes added, and one offshoot subtracted. Depending on the relative weighting of gaps, offshoots and multiple head subjects, either lifting can minimize the total misfit. In fact, the gaps and offshoots are determined by the head subjects specified in a lift.

Altogether, the set of subject clusters, their head subjects, offshoots and gaps constitutes what can be referred to as a profile of the organization in consideration. Such a representation can be easily accessed and expressed as an aggregate. It can be further elaborated by highlighting representation subjects in which the organization members have been especially successful (i.e., publication in best journals, award or other recog-

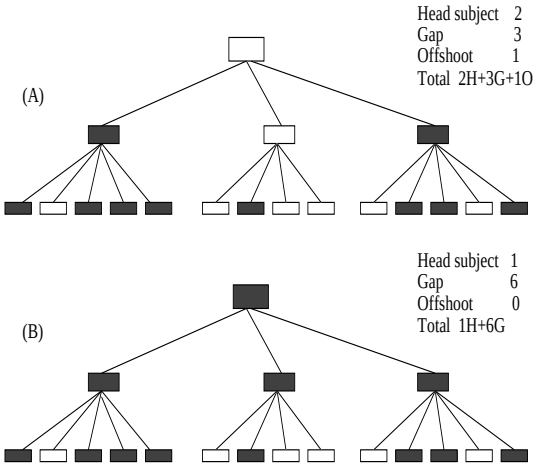

Fig. 4. Tradeoff between different liftings of the same subject cluster: mapping (B) is better than (A) if gaps are much cheaper than additional head subjects.

nition) or distinguished by another feature (i.e., industrial product or inclusion to a teaching program).

Building a parsimonious lifting of a subject cluster can be achieved by recursively building a parsimonious scenario for each node of the ACMC tree based on parsimonious scenarios for its children. At each node of the tree, sets of head gain, gap and offshoot events are to be determined and iteratively raised to the parents under each of two different assumptions that specify the situation "above the parent" starting, in fact, from the root.

One assumption is that the head subject is not at the parental node to the parent, but is somewhere higher, and the second assumption is that it has been gained in the node only. It is necessary to distinguish these two cases since, clearly, it is only meaningful to consider the loss of a head subject at a node if it was inherited at that node; similarly, it is only meaningful to consider the gain of a head if it was not inherited from above. Consider the parent-children system as shown in Figure 5, with each node assigned with sets of offshoot, gap and head gain events under the above two inheritance of head subject assumptions.

Let us denote the total number of events under the inheritance and non-inheritance assumptions by e_i and e_n, respectively. A lifting result at a given node is defined by a triplet of sets (H, G, O), representing the tree nodes at which events of head gains and gaps, respectively, have occurred in the subtree rooted at the node. We use (Hn, Gn, On) and (Hh, Gh, Oh) to denote lifting results under the inheritance and non-inheritance assumptions, respectively. The algorithm computes parsimonious scenarios for parental nodes according to the topology of the tree, proceeding from the leaves to the root in the manner, which is, to some extent similar to that described in [5].

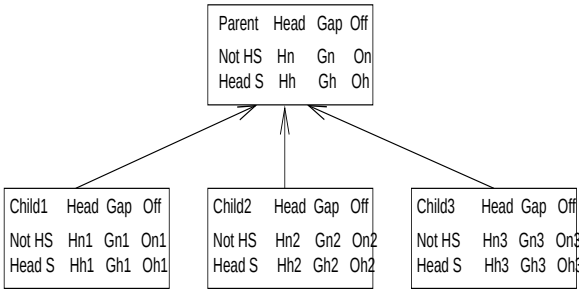

Fig. 5. Events in a parent-children system according to a parsimonious lifting scenario; HS and Head S stand for Head subject.

3 An Example of Implementation

Let us describe how this approach can be implemented by using the data from a survey conducted at the Department of Computer Science, Faculty of Science & Technology, New University of Lisboa (DI-FCT-UNL). The survey involved 49 members of the academic staff of the department.

For simplicity, we use only data of the second level of ACMC, each having a code V.v where V=A,B,...,K, and v =1,..,mK, with mK being the number of second level topics. Each member of the department supplied three ACMC topics most relevant to their current research. These comprise altogether 26 of the 59 topics at the second level in ACMC (we omit two subjects of the second level, General and Miscellaneous, occurred in every first-level division as they do not contribute to the representation). The similarity between two ACMC subjects, V.v and W.w, was defined as the number of members of the department that work on both of them.

With the algorithm ADDI-S applied to the 26x26 similarity matrix, we get the following 6 clusters (each of them contributes more than 4% to the data scatter): **Cl1** (contribution 27.08%, intensity 2.17), 4 items: D3, F1, F3, F4; **Cl2** (contribution 17.34%, intensity 0.52), 12 items: C2, D1, D2, D3, D4, F3, F4, H2, H3, H5, I2, I6; **Cl3** (contribution 5.13%, intensity 1.33), 3 items: C1, C2, C3; **Cl4** (contribution 4.42%, intensity 0.36) , 9 items: F4, G1, H2, I2, I3, I4, I5, I6, I7; **Cl5** (contribution 4.03%, intensity 0.65), 5 items: E1, F2, H2, H3, H4; **Cl6** (contribution 4.00%, intensity 0.64), 5 items: C4, D1, D2, D4, K6. These clusters lifted in the ACMC are presented on Figure 6, in which only those first-level categories that overlap them are shown.
One can see the following:

- The department covers, with a few gaps and offshoots, six head subjects shown on the Figure using pentagons filled in by different patterns;
- The most contributing cluster, with the head subject F. Theory of computation, comprises a very tight group of a few second level topics;
- The next contributing cluster has not one but two head subjects, D and H, and offshoots to every other head subject in the department, which shows that this cluster currently is the structure underlying the unity of the department;

– Moreover, the two head subjects of this cluster come on top of two other subject clusters, each pertaining to just one of the head subjects, D. Software or H. Information Systems. This means that the two-headed cluster signifies a new direction in Computer Sciences, combining D and H into a single new direction, which seems a feature of the current developments indeed; this should eventually get reflected in an update of the ACM classification (probably by raising D.2 Software Engineering to the level 1?);
– There are only three offshoots outside the department's head subjects: E1. Data structures from H. Information Systems, G1. Numerical Analysis from I. Computing Methodologies, and K6. Management of Computing and Information Systems from D. Software. All three seem natural and should be reflected in the list of collateral links between different parts of the classification tree.

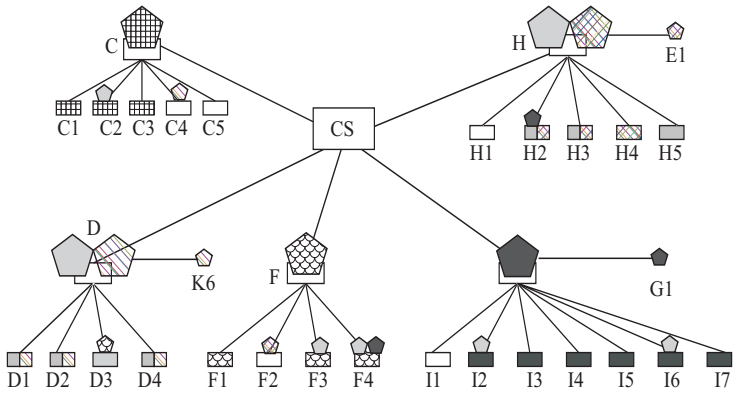

Fig. 6. Six subject clusters in the DI-FCT-UNL represented over the ACMC ontology. Head subjects are shown with differently patterned pentagons. Topic boxes shared by different clusters are split-patterned.

4 Conclusion

We have shown that ACMC can be used as an ontology structure for representing CS research activities. In principle, the approach can be extended to other areas of science or engineering, provided that these areas have been systematized into comprehensive ontologies or taxonomies. Potentially, this approach could lead to a useful instrument of visually feasible comprehensive representation of developments in any field of human activities prearranged as a hierarchy of relevant topics.

Acknowledgments The authors acknowledge all DI-FCT-UNL members that agreed to participate in the survey. Igor Guerreiro is acknowledge for the development of the e-survey tool illustrated in Figure 1. This work is supported by the grant PTDC/EIA/69988/2006 from the Portuguese Foundation for Science & Technology.

References

1. *The ACM Computing Classification System*, http://www.acm.org/class/1998/ccs98.html, 1998.
2. R. Spence, *Information Visualization*, Addison-Wesley (ACM Press), 2000.
3. B. Mirkin, *Clustering for Data Mining: A Data Recovery Approach*, Chapman & Hall /CRC Press, 2005.
4. S. Nascimento, B. Mirkin and F. Moura-Pires, Modeling Proportional Membership in Fuzzy Clustering, *IEEE Transactions on Fuzzy Systems*, **11**(2), pp. 173-186, 2003.
5. B. Mirkin, T. Fenner, M. Galperin and E. Koonin, 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.
6. G. Gallo, M.D. Grigoriadis and R.E. Tarjan, A fast parametric maximum flow algorithm and applications, *SIAM Journal on Computing*, **18**, pp. 30-55, 1989.
7. J. Shi and J. Malik, Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), pp. 888-905, 2000.