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

Boris Mirkin[1],[2], Susana Nascimento[3], Trevor Fenner[1], and Luís Moniz Pereira[3]

[1] School of Computer Science, Birkbeck University of London, London, WC1E 7HX, UK,
[2] Division of Applied Mathematics, Higher School of Economics, Moscow, RF
[3] Computer Science Department and Centre for Artificial Intelligence (CENTRIA), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

Abstract. A hybrid of two novel methods - additive fuzzy spectral clustering and lifting method over a taxonomy - is applied to analyse the research activities of a department. To be specific, we concentrate on the Computer Sciences area represented by the ACM Computing Classification System (ACM-CCS), but the approach is applicable also to other taxonomies. Clusters of the taxonomy subjects are extracted using an original additive spectral clustering method involving a number of model-based stopping conditions. The clusters are parsimoniously lifted then to higher ranks of the taxonomy by minimizing the count of "head subjects" 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 [5] for interpretation and annotation of various gene sets and gene expression data is becoming a matter of routine in bioinformatics (see, for example, [13] 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, utilizing 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 determined by a special parsimonious “lift” procedure. 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 [10].

We apply these constructions to visualize activities of Computer Science research organizations. 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 that. An earlier stage of this project is described in [11]. This work extends the additive clustering model described in [11] to fuzzy additive clustering by using the spectral decomposition of our similarity matrix and extracting clusters one by one, which allows to draw a number of model-based stopping conditions. Found clusters are lifted to higher ranks in the ACM taxonomy and visualised via a novel recursive parsimonious lift method.

The rest of the paper is organized as follows: Section 2 introduces taxonomy-based profiles, Section 3 describes a new spectral fuzzy clustering method for deriving fuzzy clusters from the profiles, Section 4 introduces our parsimonious lift method to generalize found clusters to higher ranks in a taxonomy tree. Section 5 presents the application of the proposed cluster-lift method to a real world case.

2 Taxonomy-based profiles: why 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 [16]; a device for determining the semantic similarity in information retrieval [8] and e-learning applications [17, 4]; and a device for matching software practitioners' needs and software researchers' activities [3].

Here we concentrate on a different application of ACM-CCS – a generalized representation of a Computer Science research organization which can be utilized for analysis and planning purposes. Obviously, an ACM-CCS visualized image 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.

In our work, 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. To tackle this, and also to allow for expert evaluations, we developed an interactive E-Survey tool of Scientific Activities-ESSA (available at https://copsro.di.fct.unl.pt/), 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. 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.

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 activity profiles 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.

There is a specific branch of clustering applied to similarity data, the so-called relational fuzzy clustering [2]. In the framework of 'usage profiling' relational fuzzy clustering is a widely used technique for discovering different interests and trends among users from Web log records. In bioinformatics several clustering techniques have been successfully applied in the analysis of gene expression profiles and gene function prediction by incorporating gene ontology information into clustering algorithms [6].

In spite of the fact that many fuzzy clustering algorithms have been developed already, the situation remains rather uncertain because they all involve manually specified parameters such as the number of clusters or threshold of similarity, which are subject to arbitrary choices. This is why we come up with a version of approximate clustering modified with the spectral clustering approach to make use of the Laplacian data transformation, which has proven an effective tool to sharpen the cluster structure hidden in the data [7]. This combination leads to a number of model-based parameters as aids for choosing the right number of clusters.

Consider a set of V individuals ($v = 1, 2, \cdots, V$), 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 ($v = 1, 2, \cdots, V$).

Then the similarity $w_{tt'}$ between topics t and t' can be defined as the inner product of profiles $f_t = (f_{tv})$ and $f_{t'} = (f_{t'v})$, $v = 1, 2, \cdots, V$, modified by individual weights.

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 [7]. 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 [15, 7]. Of comparative properties of these two normalizations, the normalized Laplacian, in general, is considered superior [7].

Yet the Laplacian normalization by itself cannot be used in our approach below because our method relies on maximum rather than minimum eigenvalues. To pass over this issue, the authors of [12] utilized a complementary matrix $M_n = D^{-1/2} W D^{-1/2}$ which relates to L_n with equation $L_n = I - M_n$ where I is the identity matrix. This means that M_n has the same eigenvectors as L_n, whereas the respective eigenvalues relate to each other as λ and $1 - \lambda$, so that the matrix M_n can be utilized for our purposes as well. We prefer using 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 Λ 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 λ_1 of L_n, corresponding to the same eigenvector. The inverse of a λ near 0 could make the value of $1/\lambda$ quite large and greatly separate it from the inverses of other near zero eigenvalues of L_n. Therefore, the Lapin transformation generates larger gaps between the eigen-values of interest, which suits our one-by-one approach described in the next section well.

Thus, we utilize further on the transformed similarity matrix $A = L_n^-(W)$.

3.2 Additive fuzzy clusters using a spectral method

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 μu that is a solution rather than its individual co-factors.

Our additive fuzzy clustering model involves K fuzzy clusters that reproduce the pseudo-inverted Laplacian similarities $a_{tt'}$ up to additive errors according to the following equations:

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

where $u_k = (u_{kt})$ is the membership vector of cluster k, and μ_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 μ^2 summarizes the contribution of intensity and will be referred to as the cluster's weight.

To fit the model in (1), 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. On the other hand, if vectors u_k are constrained to be just 1/0 binary vectors, the model (1) becomes of the so-called additive clustering [14, 9].

We apply the one-by-one principal component analysis strategy for finding one cluster at a time. Specifically, at each step, we consider the problem of minimization of a reduced to one fuzzy cluster least-squares criterion

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

with respect to unknown positive ξ weight (so that the intensity μ is the square root of ξ) 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 (1)), so that the residual similarity matrix for obtaining the next cluster will be $B - \mu^2 u u^T$ where μ and u are the intensity and membership vector of the found cluster. In this way, A indeed is additively decomposed according to formula (1) 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 ξ minimizing criterion (2) 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 ξ is

$$\xi = \frac{\mathbf{u}'B\mathbf{u}}{(\mathbf{u}'\mathbf{u})^2} \tag{3}$$

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

By putting this ξ in equation (2), we arrive at

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

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

Let us denote the last item by

$$G(u) = \xi^2 \left(\mathbf{u}'\mathbf{u}\right)^2 = \left(\frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'\mathbf{u}}\right)^2, \tag{4}$$

so that the similarity data scatter is the sum:

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

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

An optimal cluster, according to (5), is to maximize the explained part $G(u)$ in (4) or its square root

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

which is the celebrated Rayleigh 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 λ and corresponding normed eigenvector z for B, $[\lambda, z] = \Lambda(B)$, and take its projection to the set of admissible fuzzy membership vectors.

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

1. The optimal value of ξ (3) 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 ϵ value, say less than 5% of the original similarity data scatter.

More details on the method referred to as ADDI–FS, including its application to affinity and community structure data along with experimental comparisons with other fuzzy clustering methods, are described in [10].

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. 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 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).

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.

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 × 46 similarity matrix, two clusters have been sequentially extracted, after which the residual matrix has become definite negative (stopping condition (1)). The contributions of two clusters total to about 50%. It should be noted that the contributions reflect the tightness of the clusters rather than their priority stand. 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 1 shows the representation of CENTRIA’s cluster 2 in the ACM-CCS taxonomy with penalties of $h = 1$, $o = 0.8$, and $g = 0.1$. Increasing the gap penalty g from 0.1 to 0.2 would lead to a different parsimonious generalization in which the head subject ’G’ is decomposed into a number of head subjects on lower ranks of the hierarchy.

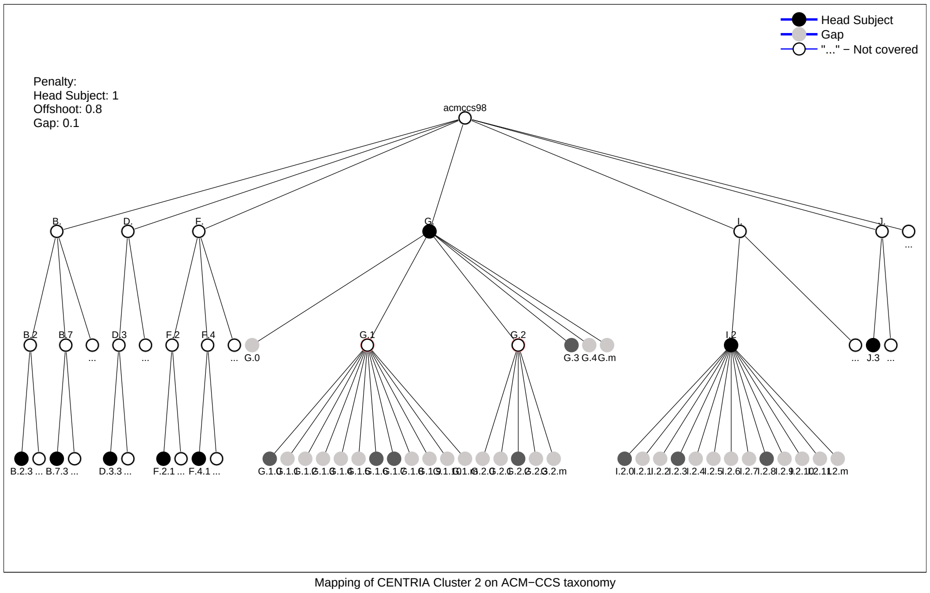

Fig. 1. Mapping of CENTRIA cluster 2 onto the ACM-CCS tree with penalties $h = 1$, $o = 0.8$ and $g = 0.1$.

6 Conclusion

We have described 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. 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.

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

The research described in this paper raises a number of issues related to all main aspects of the project: data collection, thematic clustering and lifting. On the data collection side, the mainly manual e-survey ESSA tool should be supported by an automated analysis and rating of relevant research documents including those on the internet. The ADDI-FS method, although already experimentally proven competitive to a number of existing methods, should be further explored and more thoroughly investigated. The issue of defining right penalty weights for cluster lifting should be addressed. Moreover, further investigation should be carried out with respect to the extension of this approach to more complex than taxonomy, ontology structures.

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 Figure 1. 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. 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).
4. Gaevic, D., Hatala, M.: "Ontology mappings to improve learning resource search", *British Journal of Educational Technology*, **37**(3), pp. 375 - 389 (2006).
5. "The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology", *Nature Genetics*, 25, pp. 25-29 (2000).
6. 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) doi: 10.1109/CSB.2004.1332456.
7. von Luxburg, U.: A tutorial on spectral clustering, *Statistics and Computing* 17, pp. 395-416 (2007).
8. 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) doi: 10.1109/IAT.2005.50.
9. Mirkin, B.: "Additive clustering and qualitative factor analysis methods for similarity matrices", *Journal of Classification*, **4**(1), pp. 7-31 (1987) doi:10.1007/BF01890073.
10. 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).

11. Mirkin, B., Nascimento, S., Pereira, L.M.: “Cluster-lift method for mapping research activities over a concept tree”, In: J. Koronacki, S.T. Wierzchon, Z.W. Ras, J. Kacprzyk (eds.), *Recent Advances in Machine Learning II, Computational Intelligence Series* Vol. 263, Springer, pp. 245-258 (2010).
12. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: analysis and an algorithm, In: T.G. Ditterich, S. Becker, Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems*, 14, MIT Press, Cambridge Ma., pp. 849-856 (2002).
13. 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 (2009), 3(Suppl 4):S8doi:10.1186/1753-6561-3-S4-S8.
14. Shepard, R.N., Arabie, P.: “Additive clustering: representation of similarities as combinations of overlapping properties”,*Psychological Review* 86, 87-123, (1979).
15. Shi, J., Malik, J.: “Normalized cuts and image segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), pp. 888-905 (2000).
16. Thorne, C., Zhu, J., Uren, V.: “Extracting domain ontologies with CORDER”, *Tech. Reportkmi-05-14*. Open University, pp. 1-15 (2005).
17. 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), pp. 42-52 (2005).