The Korean Journal of Applied Statistics (2008)
21(6), 1007–1014

Meta Analysis of Usability Experimental Research Using New Bi-Clustering Algorithm

Kyunga Kim[1] · Wonil Hwang[2]

[1]Dept. of Statistics, Seoul National University;
[2]Dept. of Industrial and Information Systems Engineering, Soongsil University

(Received September 2008; accepted October 2008)

Abstract

Usability evaluation(UE) experiments are conducted to provide UE practitioners with guidelines for better outcomes. In UE research, significant quantities of empirical results have been accumulated in the past decades. While those results have been anticipated to integrate for producing generalized guidelines, traditional meta-analysis has limitations to combine UE empirical results that often show considerable heterogeneity. In this study, a new data mining method called weighted bi-clustering(WBC) was proposed to partition heterogeneous studies into homogeneous subsets. We applied the WBC to UE empirical results and identified two homogeneous subsets, each of which can be meta-analyzed. In addition, interactions between experimental conditions and UE methods were hypothesized based on the resulting partition and some interactions were confirmed via statistical tests.

Keywords: Data mining, meta-analysis, clustering, usability evaluation.

1. Introduction

Usability evaluation(UE) is an important research area in Human-Computer Interaction(HCI) because it can direct optimal ways to improve the quality of systems or products in terms of ease of use. In the past twenty years, various UE methods have been developed, including think aloud (Lewis, 1982), heuristic evaluation (Nielsen and Molich, 1990) and cognitive walkthrough (Polson *et al.*, 1992) and employed for UE experiments. Respective UE studies provided only context-based conclusions(*e.g.*, usability inspection results based on a specific task scenario performed by graduate students as evaluators). More generalized conclusions are necessary for investigating UE research issues, such as task and evaluator effects (Lewis, 2001) and for providing meaningful knowledge and guidelines to usability practitioners. Because a significant quantity of empirical results have been accumulated from previous UE research, generalized guidelines derived from the combined

This work was supported by the Soongsil University Research Fund and the National Research Laboratory Program of Korea Science and Engineering Foundation(M10500000126).
[1]Postdoctoral Researcher, Dept. of Statistics, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea. E-mail: kyunga.j.kim@gmail.com
[2]Corresponding author: Professor, Dept. of Industrial and Information Systems Engineering, Soongsil University, 511 Sangdo-dong, Dongjak-gu, Seoul 156-743, Korea. E-mail: wonil@ssu.ac.kr

results can be as useful as new research that involves all the factors. However, there have been few efforts to synthesize the empirical results of usability evaluation research and to develop new useful implications for usability practitioners from the body of studies.

While meta-analysis approach is the most common method to integrate the quantitative results of previous empirical research, it has been not applicable to a comparison study of UE methods due to two reasons. First, there were not enough descriptive statistics in the HCI literature (Hartson *et al.*, 2003). Second, the experimental conditions are likely to be heterogeneous among UE studies. Therefore, alternative methods to traditional meta-analysis need to be developed for combining results of previous UE studies to draw new implications. Data mining is the methodology of analyzing large amounts of observational data to find useful novel models or patterns (Hand *et al.*, 2001). Clustering is one of the popular methods in data mining, which tries to describe the data set via data segmentation. In a cluster analysis, all of the data is segmented into subsets or clusters, such that data within a cluster are more similar to one another than those assigned to different clusters (Johnson and Wichern, 2002). Therefore, clustering can be used to partition empirical results into rather homogeneous subsets.

Most clustering algorithms, such as hierarchical clustering, K-means clustering and Gaussian mixture model, are used either to construct clusters of cases or clusters of variables in one-way. Alternatively, there has been another clustering approach, called bi-clustering(BC) or two-way clustering, that clusters cases and variables simultaneously. A major advantage of BC is that resulting clusters are interpreted directly on the data matrix expressing the interaction between the two marginal clusters. Recently various BC algorithms have been proposed for the analysis of gene expression data (Cheng and Church, 2000; Lazzeroni and Owen, 2000; Tang *et al.*, 2001; Kluger *et al.*, 2003; Yang *et al.*, 2003). Based on these algorithms, whole data can be segmented into bi-clusters that consist of subsets of genes(rows) and jointly respond across a subset of conditions(columns). The BC methods are categorized according to BC models, such as additive, multiplicative and probabilistic models and their algorithms, such as iterative row and column clustering combination, divide and conquer, greedy iterative search, exhaustive bi-cluster enumeration and distribution parameter identification algorithms (Madeira and Oliveira, 2004).

While BC methods have apparent advantages, such as direct interpretation of interactions between bi-clusters and a lot of application areas(*e.g.*, information retrieval and text mining, collaborative filtering, recommendation systems and target marketing and marketing research), current BC algorithms still have room for improvement. For example, current BC algorithms assume equal weights in all data points, but the equal-weight assumption can be unrealistic in certain situations. In this study, we develop a new data mining method, called weighted bi-clustering(WBC) algorithm, which accommodates the heterogeneity among multiple studies. Therefore, WBC provides a new way to integrate the empirical results from those similar but not homogeneous studies. Our new method is employed to integrate individual knowledge of UE studies and to extract new knowledge, including the interactions that are identified from the resulting bi-clusters. Our WBC is more appropriate than BC for two folds. Because sample sizes are different across individual studies in most integrative analyses, individual studies often produce different qualities of information. In addition, previous UE studies were conducted based on specific contexts and hence their empirical results usually contain different sets of experimental conditions. Therefore, individual UE studies cannot be considered to have equal weights.

2. Weighted Bi-Clustering

The proposed WBC algorithm is a residue-based algorithm and an extension of Cheng and Church's (2000) BC algorithm which was based on an additive model and a node-deletion procedure. In this study, the cases were replaced by the individual experiments, whose results cannot be treated in the same way because of their different sample sizes. Thus, the individual experiments were weighted by a weight matrix in addition to a data matrix in a new WBC algorithm using their sample sizes.

2.1. BC vs. WBC

The basic idea of BC came from Hartigan (1972), who explained that two-way clustering of a data matrix is better than two one-way clusterings. The BC algorithms deal with the problem of grouping cases and variables simultaneously. One commonly used BC algorithm is based on additive model and a node-deletion algorithm (Cheng and Church, 2000; Madeira and Oliveira, 2004). The additive model assumes that the elements of bi-clusters can be predicted by the sum of row effect and column effect. Note that the interaction effect of row and column was not considered in Cheng and Church's method as in other BC methods with additive models. With this additive model, Cheng and Church's node-deletion algorithm defined a bi-cluster as a subset of rows and columns with a high-similarity score. In their algorithm, the mean squared residue was introduced as a similarity score index, which is a measure of the coherence of the rows and columns in the bi-cluster and a δ-bi-cluster was defined as a submatrix that has its mean squared residue below δ for some $\delta \geq 0$. Cheng and Church assumed that all rows and columns are given equal weights in the computation of the mean squared residue, even though they indicated that there might be some doubts as to why the same weights are given to the rows and columns.

In this study, in order to overcome the equal weight assumption, the node-deletion algorithm for WBC was developed. In addition, WBC provides an alternative way to deal with data matrices containing missing values, by constructing an appropriate weight system. Cheng and Church's BC algorithm utilized the random masking method that replaces both missing values and bi-clusters with random numbers. The random masking is criticized by the random interference phenomenon, in which there might be a substantial risk that random numbers will interfere with the discovery of next bi-clusters (Yang *et al.*, 2003). This problem could be prevented by using proper weights in the node-deletion algorithm.

2.2. Additive model and weighted mean squared residue

Let A and a_{ij} denote a data matrix and its $(i, j)^{th}$ element. The additive model with weights is as follows:

$$a_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where μ is the background effect; α_i is the i^{th} row effect; β_j is the j^{th} column effect; ϵ_{ij} is the random error and independently distributed as $N(0, w_{ij}^{-1}\sigma^2)$; and w_{ij} is the weight corresponding to the $(i, j)^{th}$ element. Compared to the additive model with equal weights, this additive model replaces the strict assumption of equal variances of random errors with less strict assumption that variance of random errors is different according to the unequal weights. In order to assess the coherence of the rows and columns in the bi-cluster, we developed the weighted mean squared residue(WMSR) as a weighted version of the mean squared residue(MSR) that was used in Cheng and Church (2000)'s

algorithm. Denote a submatrix of A and its corresponding weight sums as below:

$$A_{IJ} = (a_{ij})_{i\in I, j\in J}, \qquad w_{IJ} = \sum_{i\in I, j\in J} w_{ij}, \qquad w_{iJ} = \sum_{j\in J} w_{ij}, \qquad w_{Ij} = \sum_{i\in I} w_{ij}.$$

Assume that $\sum_{i\in I, j\in J} w_{ij} = 1$, $\sum_{i\in I} w_{ij}\alpha_i = 0$ for each j and $\sum_{j\in J} w_{ij}\beta_j = 0$ for each i. The weighted grand mean and the weighted row and column means for A_{IJ} are calculated respectively as follows:

$$\begin{aligned}
\bar{a}^w_{IJ} &= \sum_{i\in I, j\in J} w_{ij}a_{ij} = \mu + \sum_{i\in I, j\in J} w_{ij}\epsilon_{ij},\\
\bar{a}^w_{iJ} &= \sum_{j\in J}\left(\frac{w_{ij}}{w_{iJ}}\right)a_{ij} = \mu + \alpha_i + \sum_{j\in J}\left(\frac{w_{ij}}{w_{iJ}}\right)\epsilon_{ij},\\
\bar{a}^w_{Ij} &= \sum_{i\in I}\left(\frac{w_{ij}}{w_{Ij}}\right)a_{ij} = \mu + \beta_j + \sum_{i\in I}\left(\frac{w_{ij}}{w_{Ij}}\right)\epsilon_{ij}.
\end{aligned}$$

Then, the estimates of all effects in the additive model are described as below:

$$\hat{\mu} = \bar{a}^w_{IJ}, \qquad \hat{\alpha}_i = \bar{a}^w_{iJ} - \hat{\mu} \text{ and } \hat{\beta}_j = \bar{a}^w_{Ij} - \hat{\mu}$$

and the weighted mean squared residue WMSR(I, J) is defined as:

$$\text{WMSR}(I, J) = \sum_{i\in I, j\in J} w_{ij}\left(a_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j\right)^2 = \sum_{i\in I, j\in J} w_{ij}\left(a_{ij} - \bar{a}^w_{iJ} - \bar{a}^w_{Ij} + \bar{a}^w_{IJ}\right)^2.$$

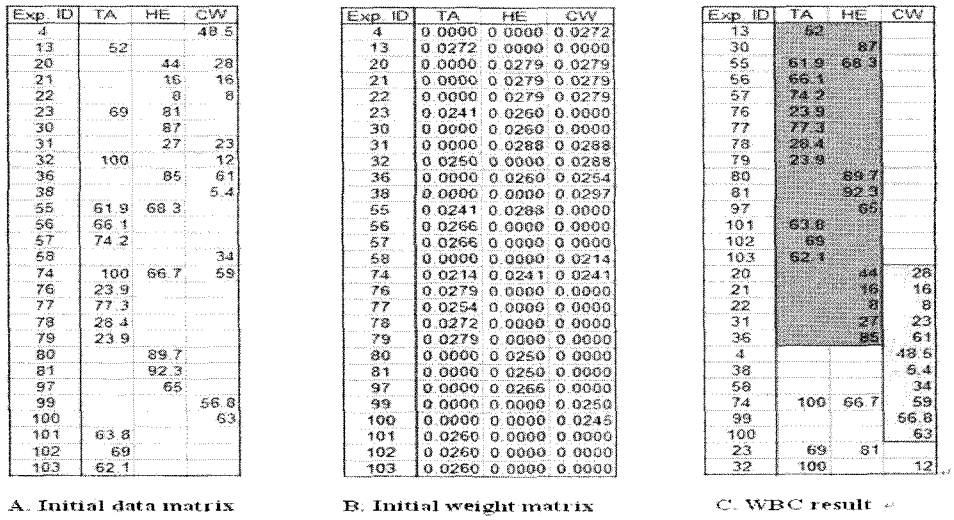
Low scores of WMSR(I, J) indicate strong coherence among the elements of the submatrix A_{IJ}, because WMSR(I, J) is a measure of the coherence of the rows and columns in the submatrix A_{IJ}.

2.3. Node-deletion algorithm

We propose a node-deletion algorithm for WBC, which is an iterative procedure to identify bi-clusters. This is a generalized version of Cheng and Church (2000)'s algorithm in that it accommodates all kind of weights. In other words, Cheng and Church's algorithm is a special case that assumes equal weights. The node-deletion algorithm consists of three main steps: WMSR calculation, deletion and update. At Step 1, WMSR(I, J) are computed for an input data matrix A_{IJ}. If WMSR(I, J) $\leq \delta^w$ for a non-negative predetermined threshold δ^w, A_{IJ} is declared as a δ^w bi-cluster and the procedure completes. Otherwise, go to Step 2 at which the contributions of each row and that of each column to the heterogeneity of the A_{IJ} are assessed via the following measures:

$$\begin{aligned}
d^w(i) &= \sum_{j\in J}\left(\frac{w_{ij}}{w_{iJ}}\right)\left(a_{ij} - \bar{a}^w_{iJ} - \bar{a}^w_{Ij} + \bar{a}^w_{IJ}\right)^2,\\
d^w(j) &= \sum_{i\in I}\left(\frac{w_{ij}}{w_{Ij}}\right)\left(a_{ij} - \bar{a}^w_{iJ} - \bar{a}^w_{Ij} + \bar{a}^w_{IJ}\right)^2.
\end{aligned}$$

Then, a row or a column with the largest contribution is detected and removed from A_{IJ}. At Step 3, the submatrix after deletion is considered as the updated input data matrix. For example, the i^{th} row was detected at Step 2, the updated input data matrix is $A_{I'J}$ with $I' = I - \{i\}$. The weights corresponding to the deleted row or column are also deleted from the weight matrix and thus the weight matrix after deletion is considered as the updated one. These three steps are iterated until a



Exp_ID	TA	HE	CW
4			48.5
13	52		
20		44	28
21		16	16
22		8	8
23	69	81	
30		87	
31		27	23
32	100		12
36		85	61
38			5.4
55	61.9	68.3	
56	66.1		
57	74.2		
58			34
74	100	66.7	59
76	23.9		
77	77.3		
78	28.4		
79	23.9		
80		89.7	
81		92.3	
97		65	
99			56.8
100			63
101	63.8		
102	69		
103	62.1		

A. Initial data matrix

Exp_ID	TA	HE	CW
4	0.0000	0.0000	0.0272
13	0.0272	0.0000	0.0000
20	0.0000	0.0279	0.0279
21	0.0000	0.0279	0.0279
22	0.0000	0.0279	0.0279
23	0.0241	0.0260	0.0000
30	0.0000	0.0260	0.0000
31	0.0000	0.0288	0.0288
32	0.0250	0.0000	0.0288
36	0.0000	0.0260	0.0254
38	0.0000	0.0000	0.0297
55	0.0241	0.0288	0.0000
56	0.0266	0.0000	0.0000
57	0.0266	0.0000	0.0000
58	0.0000	0.0000	0.0214
74	0.0214	0.0241	0.0241
76	0.0279	0.0000	0.0000
77	0.0254	0.0000	0.0000
78	0.0272	0.0000	0.0000
79	0.0279	0.0000	0.0000
80	0.0000	0.0250	0.0000
81	0.0000	0.0250	0.0000
97	0.0000	0.0266	0.0000
99	0.0000	0.0000	0.0250
100	0.0000	0.0000	0.0245
101	0.0260	0.0000	0.0000
102	0.0260	0.0000	0.0000
103	0.0260	0.0000	0.0000

B. Initial weight matrix

Exp_ID	TA	HE	CW
13	52		
30		87	
55	61.9	68.3	
56	66.1		
57	74.2		
76	23.9		
77	77.3		
78	28.4		
79	23.9		
80		89.7	
81		92.3	
97		65	
101	63.8		
102	69		
103	62.1		
20		44	28
21		16	16
22		8	8
31		27	23
36		85	61
4			48.5
38			5.4
58			34
74	100	66.7	59
99			56.8
100			63
23	69	81	
32	100		12

C. WBC result

Figure 3.1. WBC for 28 UE experiments(Note that Exp. ID denotes experiment identification number)

bi-cluster is identified. Once a bi-cluster is declared, all rows and columns in the declared bi-cluster are removed from the data matrix; and another bi-cluster is further searched in the updated data matrix.

3. Application of WBC to Usability Evaluation Research

3.1. Raw data

In this study, we collected 28 UE experiments from the previous studies accumulated in UE research to produce comprehensive knowledge on the relationship between main factors affecting the criteria of usability, such as UE methods and experimental conditions. UE methods can be classified into the expert-based and the user-based approaches and the three most widely used methods are think aloud(TA), heuristic evaluation(HE) and cognitive walkthrough(CW) (Hertzum and Jacobsen, 2001). The experimental conditions include evaluators, tasks and evaluated systems(or products). Each condition still remains as an important research issue (Lewis, 2001) and has been recognized as major limitations of individual UE studies (Andre *et al.*, 2003). The criteria for good UE methods include overall discovery rate(ODR or thoroughness), the number of usability problems detected by evaluators, satisfaction, validity, effectiveness and reliability (Hartson *et al.*, 2003).

3.2. Construction of data and weight matrices

When applying WBC to the 28 UE experiments, we used the ODRs as data values, UE methods as variables and individual experiments as cases (Figure 3.1A). Thus the data matrix consists of 28 rows and three columns. Based on the resulting bi-clusters, possible grouping factors of UE studies will be explored. For instance, we considered simultaneously individual experiments that contain the experimental conditions as hidden grouping factors and usability evaluation methods as the explicit grouping factors. In addition to the data matrix, we developed the weight matrix to

Table 3.1. Association between bi-clusters and candidate grouping factors

Candidate factor	Factor levels	BC Algorithm	
		Cheng and Church's	WBC
UE method	Think aloud Heuristic evaluation Cognitive walkthrough	$\chi^2(4) = 3.87$ $p = 0.4244$	$\chi^2(2) = 32.00$ $p < 0.0001$
Evaluation unit	Individual	$\chi^2(2) = 1.48$	$\chi^2(2) = 0.80$
	Team	$p = 0.4778$	$p = 0.3713$
Evaluator's expertise	Novice, Domain expert	$\chi^2(6) = 5.24$	$\chi^2(3) = 0.40$
	HCI expert, Double expert	$p = 0.5130$	$p = 0.9403$
Evaluated-system fidelity	Low	$\chi^2(2) = 0.34$	$\chi^2(1) = 2.49$
	High	$p = 0.8436$	$p = 0.1146$
Task type	Free exploration	$\chi^2(2) = 3.68$	$\chi^2(1) = 3.87$
	Task scenario	$p = 0.1587$	$p = 0.0492$
Time constraint	No	$\chi^2(2) = 1.86$	$\chi^2(1) = 1.95$
	Yes	$p = 0.3953$	$p = 0.1628$
Report type	Video taping/observing Free style written report Structured written report	$\chi^2(4) = 7.20$ $p = 0.1257$	$\chi^2(2) = 11.31$ $p = 0.0035$

reflect the different importance of each data point. For instance, the data in meta analyses must be weighted by different sample sizes or variances with which the data comes from the summarized results of the previous experiments. In this case study, the data values are ODRs which increases as the number of the evaluators increases. In order to make the ODRs with different numbers of evaluators comparable, the ODRs with large number of evaluators need to be penalized properly, compared to the ODRs with small number of evaluators. It is widely known in UE research that the ODRs show a specific nonlinear increasing patterns, so called the asymptotic curve, as the number of evaluators increases (Hartson *et al.*, 2003). In this study, we used a least square method to find the asymptotic curve that fitted the data best and developed penalty-type weights based on this asymptotic curve. As a result, a penalty-type weight matrix (Figure 3.1B) was constructed to remove the effect of the number of evaluators on the ODRs. Note that the results from WBC can depend on the weight matrix and the weight matrix should be carefully constructed. We will demonstrate this by comparing the results from WBC with those from BC(*i.e.*, equal weights).

3.3. WBC and post-analysis of identified bi-clusters

We applied WBC to the constructed data and weight matrices. The node-deletion algorithm for WBC was conducted with several values of δ^w. We compared the results from WBC with different δ^w values and found that $\delta^w = 0.01$ resulted in two non-trivial and meaningful bi-clusters (Figure 3.1C). Because the bi-cluster sizes are maximal under the condition of WMSR ≤ 0.01, these two bi-clusters are optimal in terms of bi-cluster size and their coherence. The first bi-cluster grouped 20 experiments and two UE methods(*i.e.*, think aloud + heuristic evaluation) simultaneously and included 21 ODRs. The second bi-cluster grouped 11 experiments with one UE method(*i.e.*, cognitive walkthrough) and included 11 ODRs.

In order to investigate the quality and usefulness of the identified bi-clusters, we conducted a post-analysis on the WBC results. First, the ODRs in the first bi-cluster have overall higher values than the ODRs in the second bi-cluster(*i.e.*, two-sample one-sided t-test: $t = 2.2763$, $df =$

23.6, p-value = 0.0161). Therefore two bi-clusters seem distinct to each other and hence showed heterogeneity. The first and the second bi-clusters can represent the high-ODR group and the low-ODR group, respectively. Second, we investigated potential hidden factors that may characterize the bi-clusters. While the UE method is the explicit factor that represents a column effect, we found the experimental conditions as candidate hidden factors for UE studies. Examples of experimental conditions include units of evaluation, evaluator's expertise, fidelity of evaluated systems, task types, time constraints and report types. In order to examine those hidden factors, we further conducted chi-square tests for association between bi-clusters and candidate hidden factors. With a significance level of 5%, two hidden factors(*i.e.*, task types and report types) as well as UE methods were identified with strong association with the bi-clusters from WBC (Table 3.1).

For comparison, we also applied Cheng and Church's BC algorithm and three bi-clusters were identified. Unlike the bi-clusters from WBC, the bi-clusters from Cheng and Church's algorithm did not show significant association with any candidate factor (see Table 3.1). Therefore, WBC provided more meaningful bi-clusters than Cheng and Church's BC algorithm in this case study.

4. Discussion

Usability practitioners need to know how to optimize usability evaluation outcomes. The integrative analysis of the empirical results from previous UE studies can provide useful and generalized guidelines for practitioners. Most common integration approaches are traditional meta-analysis methods, but they may not be applicable to synthesizing the UE empirical results, which show considerable heterogeneity and unequal sample sizes. In this study, a new data mining method, called WBC, was proposed to make heterogeneous empirical results comparable via proper weights and hence to facilitate more effective integration. The WBC was applied to 38 ODRs of usability problems derived from experiments, which were reported in publications. Two distinct bi-clusters were identified and successfully characterized as high and low ODR groups. Also, potential hidden grouping factors were found based on the resulting bi-clusters. Some hidden factors, such as task types and report types, were confirmed by the statistical test for the association between candidate grouping factors and the identified bi-clusters. The interactions between grouping factors, including UE methods and experimental conditions, can be further investigated to produce guidelines for designing future UE experiments(*e.g.*, when a certain UE method is used, which type of tasks produce better usability evaluation outcomes?).

In the newly developed node-deletion algorithm for WBC, the weighted mean squared residue(WMSR) score represents how homogeneous or coherent the data within the bi-cluster are(*i.e.*, the smaller WMSR is, the more coherent the data within a bi-cluster are) and threshold δ^w is a criterion when we may accept the coherence level of bi-clusters. For example, small δ^w leads to many small-sized bi-clusters because the node-deletion algorithm tends to produce the small size of bi-clusters to reach the small WMSR score. Because bi-clusters of too small size are likely to be trivial cases, one wants to obtain maximal-sized bi-clusters with the acceptable levels of coherence. The choice of δ^w is crucial for the success of WBC and will be investigated via simulations in the future. In our case study, we demonstrated that WBC can provide more meaningful results than Cheng and Church (2000)'s BC. However, the performance of WBC need to be also further validated and compared with other BC methods via simulations.

References

Andre, T. S., Hartson, H. R. and Williges, R. C. (2003). Determining the effectiveness of the usability problem inspector: A theory-based model and tool for finding usability problems, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **45**, 455–482.

Cheng, Y. and Church, G. M. (2000). Bi-clustering of expression data, In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, 93–103.

Hand, D. J., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining*, The MIT Press, Massachusetts.

Hartigan, J. A. (1972). Direct clustering of a data matrix, *Journal of the American Statistical Association*, **67**, 123–129.

Hartson, H. R., Andre, T. S. and Williges, R. C. (2003). Criteria for evaluating usability evaluation methods, *International Journal of Human-Computer Interaction*, **15**, 145–181.

Hertzum, M. and Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods, *International Journal of Human-Computer Interaction*, **13**, 421–443.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall, New Jersey .

Kluger, Y., Basri, R., Chang, J. T. and Gerstein, M. (2003). Spectral biclustering of microarray data: Co-clustering genes and conditions, *Genome Research*, **13**, 703–716.

Lazzeroni, L. and Owen, A. (2000). Plaid models for gene expression data, *Technical Report*, Stanford University.

Lewis, C. (1982). Using the 'thinking-aloud' method in cognitive interface design, *Research Report RC9265*, IBM T. J. Watson Research Center, New York.

Lewis, J. R. (2001). Introduction: Current issues in usability evaluation, *International Journal of Human-Computer Interaction*, **13**, 343–349.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey, *IEEE Transactions on Computational Biology and Bioinformatics*, **1**, 24–45.

Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interface, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People*, 249–256.

Polson, P. G., Lewis, C., Rieman, J. and Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces, *International Journal of Man-Machine Studies*, **36**, 741–773.

Tang, C., Zhang, L., Zhang, A. and Ramanathan, M. (2001). Interrelated two-way clustering: An unsupervised approach for gene expression data analysis, In *Proceedings of Second IEEE International Symposium on Bioinformatics and Bioengineering*, 41–48.

Yang, J., Wang, H., Wang, W. and Yu, P. (2003). Enhanced bi-clustering on expression data, In *Proceedings of Third IEEE Conference on Bioinformatics and Bioengineering*, 321–327.