

Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

Title	Multi-view clustering for mining heterogeneous social network data
Author(s)	Greene, Derek; Cunningham, Pádraig
Publication Date	2009-03
This item's record/more information	<a href="http://hdl.handle.net/10197/1891">http://hdl.handle.net/10197/1891</a>

Downloaded 2012-05-16T20:40:54Z

Some rights reserved. For more information, please see the item record link above.



# Multi-View Clustering for Mining Heterogeneous Social Network Data\*

Derek Greene, Pádraig Cunningham  
University College Dublin

March, 2009

## Abstract

Uncovering community structure is a core challenge in social network analysis. This is a significant challenge for large networks where there is a single type of relation in the network (*e.g. friend* or *knows*). In practice there may be other types of relation, for instance demographic or geographic information, that also reveal network structure. Uncovering structure in such multi-relational networks presents a greater challenge due to the difficulty of integrating information from different, often discordant views. In this paper we describe a system for performing cluster analysis on heterogeneous multi-view data, and present an analysis of the research themes in a bibliographic literature network, based on the integration of both co-citation links and text similarity relationships between papers in the network.

## 1 Introduction

The challenge of integrating different perspectives on a problem in order to offer a more complete picture arises in a variety of contexts. In this paper we focus on the problem of exploring multiple associated views on a social network. Specifically we consider a bibliographic network analysis task, where research papers can be clustered based on both co-citation relationships and abstract text similarity to reveal active research themes [7]. This analysis of the case-based reasoning (CBR) conference literature network is described in Section 4.

For some data exploration applications, we may have access to a set of views that are entirely *compatible* – the same patterns will occur across all views. The problem then becomes the identification of a single consensus model describing the patterns common to the views [2]. However, in many real-world data integration scenarios there can be a significant degree of discord between the patterns present in different views. For instance in the analysis of the research literature, it is in the nature of co-citation relationships that links do not reveal themselves for a few years until publications begin to accrue citations. A co-citation link exists between two papers when these papers are both cited by a third paper. It has been shown that co-citation links provide better evidence of thematic structure than direct citation links [12]. However, since co-citation structure takes some years to develop, there will be structure in evidence based on text similarity that is not evident in the co-citation network. By the same token, in the CBR literature analysed here, there are some clusters evident in the co-citation view that are not supported by the text view. This is presumably because the researchers publishing in this research area use a diverse vocabulary. Another important aspect of many real-world

---

\*The work was part supported by Science Foundation Ireland Grant Nos. 05/IN.1/I24 and 08/SRC/I1407

data integration tasks is that the available data sources will often be *incomplete* in nature. (*i.e.* each view may only represent a subset of data objects in the domain). For instance, no citation information whatsoever may be available for certain papers or authors in a bibliographic network.

The problem of reconciling discordant models from different views has recently been referred to as learning in “parallel universes” (PU) by Berthold & Patterson [1]. The framework they propose involves the idea of sharing information between views in order to construct a set of *local models* for those views, which are subsequently combined to produce a more comprehensive *global model* of the patterns present in the domain. From a practical perspective, a key aspect of the PU framework is that it supports integration problems where structures exist in some views but not in others.

In this paper we discuss the Parallel Integration Clustering Algorithm (PICA), an approach based on the PU framework for aggregating information from heterogeneous, incomplete, and potentially discordant views. PICA was initially applied in the context of bioinformatics to help identify functional from diverse biological data sources [5]. In this paper we show how PICA can be used to explore multi-relational data in social network analysis tasks. In addition to the clustering algorithm itself, we present the *PICA Browser* application, a new data exploration tool which supports the explanation and visualisation of the clusterings produced by PICA.

In the next section we provide more detail on the motivating scenarios for unsupervised learning from multiple views. In Section 3 we discuss the operation of PICA, our proposed approach for multi-view clustering that addresses many of the challenges raised by real-life integration tasks. An evaluation of the operation of PICA on the CBR conference literature data is presented in Section 4. The paper concludes with a summary and an outline of our plans for future work.

## 2 Motivating Problems

In [7] we presented an analysis of the research themes in the field of case-based reasoning (CBR), which involved examining publication co-citation links in the research literature. The analysis was based on a core set of 672 papers, with co-citation data coming from a set of 3461 papers that cite these papers. While co-citation analysis has proven to be effective at uncovering relational structure in the research literature, it has the shortcoming that recent papers will have few co-citation links as papers citing pairs of papers in the core set have not yet appeared.

In this paper we show that PICA can be used to integrate a new source of information, based on the similarity of paper abstracts, with the co-citation data in order to provide a more comprehensive view of the research themes in the CBR literature. Our evaluation shows that PICA meets this objective of bringing recently published papers into the clustering process by incorporating the text similarity view of the data. The text view also allows older papers that did not attract a large number of citations (and thus do not have many co-citation links) to be incorporated into the final clustering. Whether this is always desirable is debatable, and it raises interesting questions about the management of the contributions from different views when learning from multiple sources.

## 3 PICA

In this section we provide a detailed description of PICA, the system we propose for integrating two or more heterogeneous data sources, using an approach based on cumulative voting in

unsupervised ensembles [3]. This algorithm can be regarded as sharing certain similarities with *late integration* multi-view classification techniques [9], as it seeks to combine previously generated clusterings produced independently on each view. However, motivated by the PU framework [1], PICA differs from standard unsupervised fusion strategies in that it allows for the fact that patterns may be present or detectable in one view, but not in another. Rather than producing an aggregated model that focuses solely on patterns common to all views, PICA conserves those that are present in a subset of the available views.

### 3.1 Algorithm Overview

Firstly we formally describe the data integration task. Let  $\mathcal{X}$  denote the set of all possible data objects in a domain of interest (*e.g.* the nodes of a social network). In this domain, we have access to a set of  $v$  views, where  $\mathcal{X}_l \subseteq \mathcal{X}$  denotes the subset of objects present in the  $l$ -th view. These objects may either be represented explicitly in a feature space, or implicitly in the form of a pairwise relation-based representation.

Rather than working on the original data, PICA takes as its input a collection of “base clusterings” constructed independently on each available view. These will typically be generated by applying a partitional algorithm such as  $k$ -means that will frequently converge to different local minima. We denote the collection of clusterings generated on the view  $\mathcal{X}_l$  by  $\mathcal{C}_l$ , and the complete set of base clusterings for all views by  $\mathbb{C} = \{\mathcal{C}_1 \cup \dots \cup \mathcal{C}_v\}$ . Given the input  $\mathbb{C}$ , PICA follows a two-stage process:

1. Produce a set of *local models*  $\{L_1, \dots, L_v\}$ , where  $L_l$  represents a model, in the form of a “soft” clustering (*i.e.* a clustering with non-negative real-valued membership weights that allows the representation of overlaps between clusters), produced on the view  $\mathcal{X}_l$ , with some contribution or “mixing” from the other views.
2. Combine the local models to produce a *global model*  $G$  (in the form of a soft clustering of all data objects in the domain). This model merges the common aspects of the local models, while preserving those clusters that are unique to each local model.

An illustration of the complete process is shown in Figure 1.

### 3.2 Local Model Construction

To initialise the local model for the view  $\mathcal{X}_l$ , we select the most representative base clustering from the set  $\mathcal{C}_l$  generated on that view, using a measure of clustering “stability”. Specifically, we calculate the *average normalised mutual information* (ANMI) [11] for each base clustering in  $\mathcal{C}_l$  with respect to the other base clusterings generated on the same view. This measures the amount of information shared between a single clustering and the other clusterings in  $\mathcal{C}_l$ . We select the base clustering from  $\mathcal{C}_l$  with the highest ANMI score as our initial local model  $L_l$ .

Next we attempt to improve our initial local model  $L_l$  by adding information from the remaining base clusterings in the complete collection  $\mathbb{C}$  that were generated on **all** views (*i.e.* not simply those in  $\mathcal{C}_l$  that were generated on  $\mathcal{X}_l$ ). This has the effect of supporting “mixing” between the views, where information provided by a base clustering from one view can inform the model constructed for another view. In practice the aggregation is performed by using a variation of the cumulative voting methods that have been previously proposed for efficiently combining an ensemble of clusterings [3]. We match the clusters in each base clustering with those in the current local model  $L_l$ . The matching procedure is performed by measuring the similarity

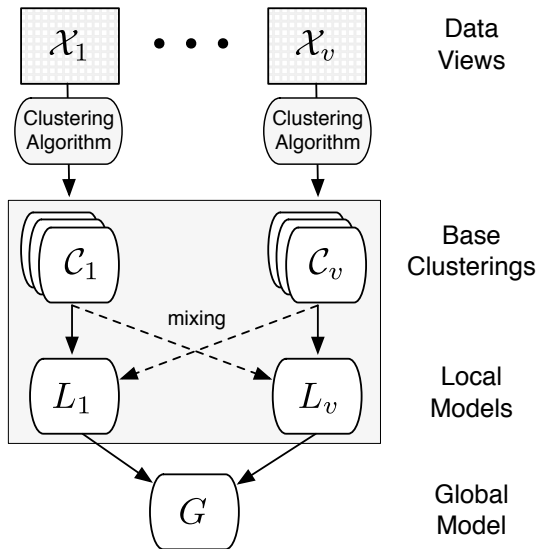


Figure 1: Overview of the Parallel Integration Clustering Algorithm (PICA).

between the clusters in two clusterings using the *binary overlap coefficient* which defines the agreement between disjoint sets  $(A, B)$  as:

$$over(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \quad (1)$$

To apply this measure to two clusterings (such as  $L_l$  and the next remaining base clustering in  $\mathbb{C}$ ), both are first converted to disjoint clusterings by thresholding, and the agreement scores between their respective disjoint clusters are calculated. The optimal correspondence between their clusters can be found by solving the minimal weight bipartite matching problem using the Hungarian method [8]. For each matched pair in the optimal correspondence, we examine the agreement score for Eqn. 1. If the score exceeds a matching threshold  $\theta \in [0, 1]$ , the pair of clusters are merged – the model  $L_l$  is updated using an average voting scheme as described in [3]. Unlike previous cumulative voting ensemble techniques, the use of a matching threshold here means that not all clusters from the base clusterings will be used in the aggregated model – for instance, noisy or irrelevant base clusters will not make a contribution to  $L_l$ .

### 3.3 Global Model Construction

At this stage we have constructed a set of local models  $\{L_1, \dots, L_v\}$ , one for each view. These may be of interest in their own right, but for ease of interpretation and evaluation, we would like to combine these partial models to produce a single global model providing a more complete picture of the domain. This is achieved by performing an additional matching procedure at this stage, where similar clusters from each local model are merged, so that redundant patterns are combined, while unique patterns are preserved.

Specifically we consider each pair of clusters across all local models, and merge those pairs with an overlap coefficient value (1) of greater than the matching threshold  $\theta$ . This is equivalent to performing complete-linkage agglomerative clustering on the local model clusters, with the cut-off threshold set to  $\theta$ . This results in a single global model  $G$  produced from all  $v$  views, where the number of clusters in this model is  $|G| \leq \sum (|L_1| + \dots + |L_v|)$ . The clusters in  $G$

represent patterns that were unique to views, as well as those that were present in two or more views (*i.e.* clusters from the local models that were merged during the final matching procedure).

For each cluster in  $G$ , provenance information is available in the form of the *contribution* to that cluster from each view. For a single cluster, the contribution of the view  $\mathcal{X}_i$  to that cluster is measured as the fraction of the total sum of membership weights coming from clusters generated on that view.

PICA also provides a measure of robustness or *reliability* for each of the clusters in  $G$ . Assuming a diverse collection of base clusterings (*e.g.* generated by an algorithm using stochastic initialisation and/or subsampling), we would like to assess the degree to which patterns repeatedly appear in base clusterings across one or more views. For each cluster in a local model  $L_l$ , we count the fraction of base clusters that contributed to that cluster (*i.e.* the number of successful matches involving that cluster that exceeded the threshold  $\theta$ ). When clusters from the local models are merged during global model construction, we measure the reliability of a cluster in  $G$  as the average reliability of the local model clusters that were merged to form that cluster. Reliability scores will fall in the range  $[0, 1]$ , with a value closer to 1 being indicative of a more robust cluster.

### 3.4 Model Visualisation

To explore the models produced by PICA, including the contributions made by each view to the models, we have developed the *PICA Browser* application<sup>1</sup>. Examples of two clusters in a global model produced from the integration of two heterogeneous views are shown in Figures 2 and 3. To highlight cluster provenance, the left-hand side of each screenshot shows the list of clusters in the global model, with the blue/green bar showing the proportion of contribution coming from each view. Note that the clusters are arranged in descending order based on their reliability scores.

When one of the views under consideration is based on text data (such as the research abstracts used in the evaluation in Section 4), we can use this data as a means of summarising the content of the clusters generated by PICA for human inspection. As part of the *PICA Browser* interface, ordered lists of representative keywords are provided for each cluster (shown at the top right-hand corner of Figures 2 and 3). These keywords were automatically identified by ranking the terms for each cluster based on their Information Gain [13]. Given a cluster of papers, the ranking of terms for the cluster is performed as follows: firstly the centroid vector of the cluster is computed on the text view; subsequently, we compute the Information Gain between the cluster centroid vector and the centroid vector for the entire set of papers. Terms that are more indicative of a cluster will receive a higher score, thereby achieving a higher ranking in the list of keywords for the cluster.

## 4 Evaluation

An initial exploration of the thematic structure of the CBR conference literature, using Non-negative Matrix Factorisation (NMF), was presented in [7]. The analysis was based on co-citation links, an established technique for identifying relationships between research papers or authors. Since co-citation data has the shortcoming that it cannot identify relationships between very recent papers or between those papers that are poorly cited, we extend that analysis by incorporating another view that is based on the similarity between the text of publication titles and abstracts.

---

<sup>1</sup>The *PICA Browser* tool and a Java implementation of PICA are both available at <http://mlg.ucd.ie/pica>

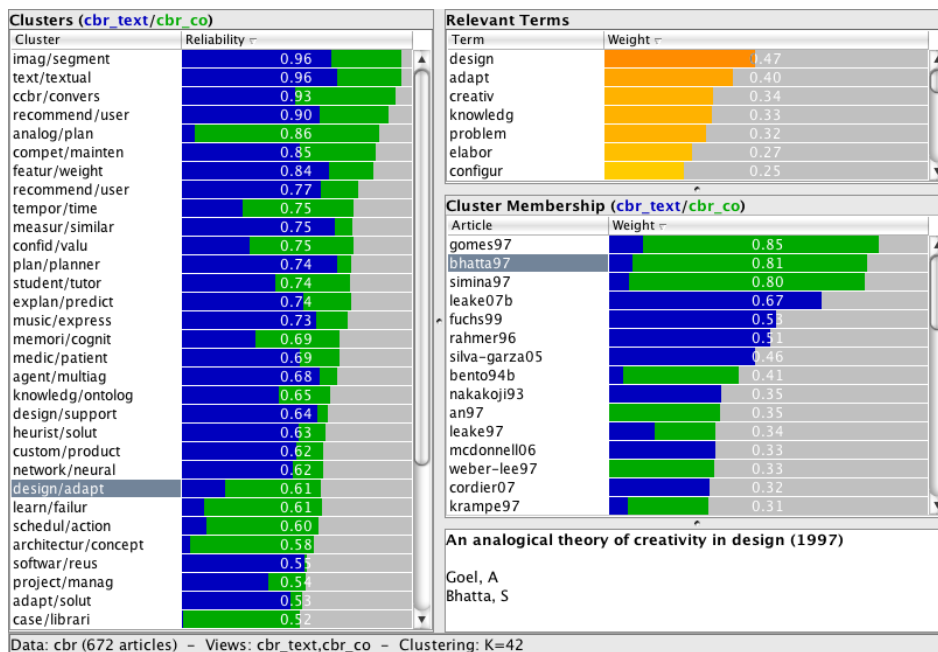


Figure 2: An example of the output of the *PICA Browser* tool. This shows a cluster of research on “adaptation”. The predominance of green (light shading) in the histogram on the right indicates that the evidence for this cluster comes mostly from the co-citation view.

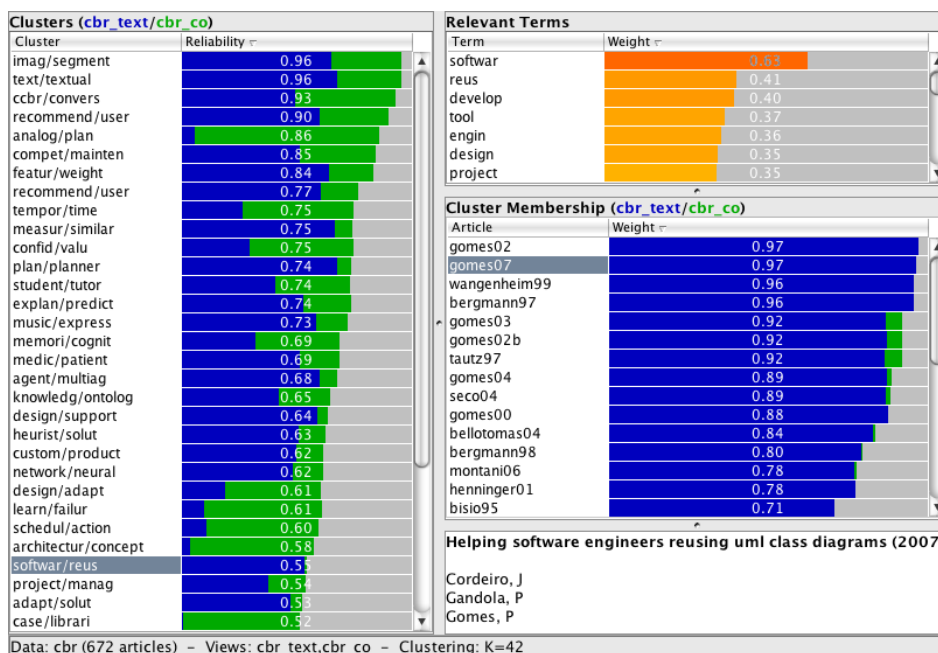


Figure 3: A cluster of research on “software reuse” shown in the *PICA Browser* tool. The predominance of blue (dark shading) in the histogram on the right shows that the support for this cluster comes from the text similarity view.

The complete CBR conference literature network dataset<sup>2</sup> consists of 672 papers published by 828 individual authors. At the time the dataset was constructed (December 2007) 518 of these papers had accrued at least one citation according to Google Scholar<sup>3</sup>, yielding an incomplete co-citation view. A text representation is available for all 672 papers, although the resulting vector space model is highly sparse, with only 1949 non-stopword terms occurring in more than one document. The goal of our evaluation was to take these two “deficient” views and use PICA to produce a superior model of the CBR research network.

## 4.1 Experimental Setup

To generate diverse collections of base clusterings on both views, the kernelised form of the  $k$ -means algorithm [10] was applied under random sub-sampling without replacement. In the case of the text data, we applied a linear kernel after normalising document term vectors to unit length. For the co-citation data we used a kernel based on the *CoCit-Score* proposed by Gmür [4] for bibliometric analysis. In both cases we shift the diagonal to zero to avoid the problems associated with diagonal dominance [6]. This was particularly important in the case of the text data, due to the sparse nature of the underlying vector space. When generating base clusterings on both views, the number of clusters was randomly chosen from  $k \in [25, 30]$  based on the analysis previously performed in [7]. To ensure robust results, a total of 2500 base clusterings was generated on each view.

When applying PICA to combine the base clusterings from the two views, we found that a matching threshold value of  $\theta = 0.3$  was most appropriate in practice. This signifies that at least 30% of one cluster must be contained in another to obtain a “reasonable” match based on the criterion given in Eqn. 1. Experiments using thresholds in the range  $\theta \in [0.2, 0.5]$  yielded highly similar clusters, with some level of duplication as  $\theta$  increased. This suggests that PICA is relatively robust to parameter changes.

## 4.2 Thematic Groupings

The fourteen research themes identified in the original study [7] that considered co-citation relations only are shown in Table 1. For the most part these themes are still evident in the clustering based on both views (text and co-citation). For instance, Figure 2 shows a cluster of papers relating to “Adaptation” that corresponds closely to one uncovered in the original analysis – the dominance of green bars in the panel on the right of the screenshot indicates that this cluster is contributed mostly by the co-citation view. However, the themes of “CBR on Temporal Problems” and “Scheduling & Agents” that were previously considered minor are now more prominent as they have strong support in the text view.

## 4.3 New Themes Revealed by PICA

We now examine four research themes revealed by the multi-view analysis that were not evident in the analysis based on co-citation only. These themes and the discriminating terms associated with them are shown in Table 2.

**CBR & Music.** This research theme covers the use of CBR in music with many of the papers having a creative or performance focus. There is some support for this theme in the co-citation

---

<sup>2</sup>Available at <http://mlg.ucd.ie/pica>

<sup>3</sup><http://scholar.google.com>

Table 1: Research themes identified in the CBR conference literature based on the co-citation view only.

<b>Major Themes</b>	
1	Recommender Systems and Diversity
2	Case-Base Maintenance
3	Case Retrieval
4	Learning Similarity Measures
5	Adaptation
6	Image Analysis
7	Textual CBR
8	Conversational CBR
9	Feature Weighting & Similarity
10	Creativity & Knowledge Intensive CBR
<b>Minor Themes</b>	
11	CBR on Temporal Problems
12	Games and Chess
13	Scheduling & Agents
14	Structural Cases

Table 2: Research themes identified in the CBR conference literature based on both the co-citation and text views, together with lists of discriminating keywords.

<b>Theme</b>	<b>Discriminating Terms</b>
CBR & Music	music, expression, perform, tempo, song, transform, phrase
Explanation	explanation, predict, explain, CBR, metric, outcome
CBR in Medicine	medicine, patient, care, health, expert, reason, therapy
Software Reuse	software, reuse, develop, tool, engineer, design, project

view but this support is not strong as many of the papers are from 2004 and later. This is a good example of the benefits of incorporating the text view, as it reveals newer research themes that are not yet supported by co-citations.

**Explanation.** This theme was already evident in the original analysis as a sub-theme of Case Retrieval – retrieving cases to support explanation is a recognised research issue in CBR. However, this theme is more evident when multi-view clustering is applied. In particular the text view helps identify a number of more recent papers from 2005 to 2007 that were not included when clustering on co-citation data alone using NMF.

**CBR in Medicine.** In the analysis based on co-citations only it was remarkable that applications of CBR in medicine did not emerge as a research theme, as this would be recognised as an application area for CBR where there is a significant amount of research activity. This theme is clearly evident in the multi-view analysis, with contributions coming from both the text and co-citation views. It may be that the reason this did not show up in the original analysis is that

much of this research is published outside the CBR conference series, and thus this theme does not have a strong signature in the available citation data.

**Software Reuse.** This cluster brings together papers on CBR for software design and design reuse that do not have much support in the co-citation view. This is unusual in that most of the papers date from 2002 to 2004, are thematically related but not connected by citations. There is also an “error” in this cluster in that it includes papers on software for CBR systems development (shown as `bergmann97`, `bellotomas04` and `bergmann98` in Figure 3), presumably because terms such as ‘software’, ‘engineering’ and ‘development’ are included in the abstract.

#### 4.4 Discussion

It is clear from this brief analysis that the incorporation of the text view brings some considerable advantages and raises some interesting questions. It seems to reinforce the structure that is evident based on co-citation and bring out some new structure. By highlighting associations with newly published papers and those papers that have not attracted many citations, we uncover a more comprehensive picture of the research themes in the CBR literature. The fact that this novel structure was not evident in the co-citation view is sometimes due to the time-lag problem with co-citation. This is not always the case however, and there are research themes revealed by text similarity that are weakly supported by links based on co-citation links (*e.g.* software reuse).

This raises the question of the status of structures that are supported from different views. In this analysis of research papers, research themes that are supported by co-citation are perhaps more important than those supported by text similarity. The mistake of linking papers on CBR tools into the cluster on CBR for software design is an example of false structure derived from the text view. Since it is important to reveal the provenance of cluster structures to the user, the *PICA Browser* has been designed to make it clear that the evidence for including these papers comes from the text view only. Our experience using the browser during the evaluation suggests that this provenance information can offer useful insights into the agreement and disagreement between structures present in related data sources.

It is worth mentioning that, in addition to identifying new clusters, the multi-view clustering in *PICA* has the added benefit of adding more recent papers to clusters that were already evident in the co-citation view. The top ranking clusters in Figures 2 and 3 were already evident in the co-citation view but now contain additional recent papers.

## 5 Conclusion

Identifying community structure is one of the core computational challenges in social network analysis. This problem is particularly difficult when there is more than one view on the data, *i.e.* there is more than one type of link between the nodes in the network. In the analysis presented here the nodes are research papers and the two types of link represent co-citation and text similarity. Because of the nature of these links, there will not always be an agreement between these two views.

We have presented *PICA*, an unsupervised data integration approach, which accommodates this disagreement by using an ensemble cumulative voting clustering framework. The examples given here show that *PICA* allows for one view to support the clustering produced by another, while also allowing for disagreement between views. In exploring the output from *PICA* it has become evident that the provenance of the cluster relationships is important. Consequently the

*PICA Browser* application has been designed with this in mind. This application highlights which views have contributed to the formation of each cluster in the global model produced by PICA, and which views influence the individual cluster assignments. We believe that this type of insight is a key requirement of any unsupervised multi-view learning system.

While the evaluation presented here suggests that the PICA strategy for integrating multiple views on social network data is effective, we wish to conduct a quantitative evaluation to further explore this. The next step in this research is to evaluate the effectiveness of PICA on annotated data to quantify the benefits of the approach.

## References

- [1] M. Berthold and D. Patterson. Towards learning in parallel universes. *Proc. 2004 IEEE International Conference on Fuzzy Systems*, 1, 2004.
- [2] S. Bickel and T. Scheffer. Multi-view clustering. In *Proc. 4th IEEE International Conference on Data Mining (ICDM'04)*, pages 19–26, Washington, DC, USA, 2004. IEEE Computer Society.
- [3] E. Dimitriadou, A. Weingessel, and K. Hornik. A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(7):901–912, 2002.
- [4] M. Gmür. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1):27–57, 2003.
- [5] D. Greene, K. Bryan, and P. Cunningham. Parallel integration of heterogeneous genome-wide data sources. In *Proc. 8th International Conference on Bioinformatics and BioEngineering (BIBE'08)*, 2008.
- [6] D. Greene and P. Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384. ACM Press, 2006.
- [7] D. Greene, J. Freyne, B. Smyth, and P. Cunningham. An analysis of research themes in the CBR conference literature. In *Proc. 9th European Conference on Case-Based Reasoning (ECCBR'08)*, pages 18–43, 2008.
- [8] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.
- [9] P. Pavlidis, J. Weston, J. Cai, and W. Noble. Learning Gene Functional Classifications from Multiple Data Types. *Journal of Computational Biology*, 9(2):401–411, 2002.
- [10] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.
- [11] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, December 2002.
- [12] H. White and C. Griffith. Author Cocitation: A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981.

- [13] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proc. 14th International Conference on Machine Learning (ICML'97)*, pages 412–420, 1997.