An Analysis of Research Themes in the CBR Conference Literature*

Derek Greene, Jill Freyne, Barry Smyth, and Pádraig Cunningham

University College Dublin
{Derek.Greene,Jill.Freyne,Barry.Smyth,Padraig.Cunningham}@ucd.ie

Abstract. After fifteen years of CBR conferences, this paper sets out to examine the themes that have evolved in CBR research as revealed by the implicit and explicit relationships between the conference papers. We have examined a number of metrics for demonstrating connections between papers and between authors and have found that a clustering based on co-citation of papers appears to produce the most meaningful organisation. We have employed an Ensemble Non-negative Matrix Factorisation (NMF) approach that produces a "soft" hierarchical clustering, where papers can belong to more than one cluster. This is useful as papers can naturally relate to more than one research area. We have produced timelines for each of these clusters that highlight influential papers and illustrate the life-cycle of research themes over the last fifteen years. The insights afforded by this analysis are presented in detail. In addition to the analysis of the sub-structure of CBR research, this paper also presents some global statistics on the CBR conference literature.

1 Introduction

To mark fifteen years of international conferences on case-based reasoning (CBR), we have set out to explore what can be learned about the internal organisation of CBR research by analysing the relationships that can be discerned from the literature. The objective is to discover the underlying themes within the literature, and to examine how these themes have evolved over the course of the conference series. A common way to perform this type of task is to apply unsupervised learning techniques to identify clusters of associated papers or authors, which correspond to thematic groups [1]. In this paper, we propose a new ensemble approach to Non-negative Matrix Factorisation (NMF) [2] for identifying such groups. We describe the application of this algorithm to the network constructed from the bibliography of the CBR conference series. From the resulting clustering, we highlight ten important research themes for discussion. We identify the influential papers within these clusters, and we also highlight those papers that have played a central role in the body of CBR literature as a whole. We hope that the results of our investigation will be of broad interest to the CBR community, as well as assisting new researchers to identify the current key themes within CBR and the seminal research papers supporting these themes.

* This research was supported by Science Foundation Ireland Grant No. 05/IN.1/I24.

K.-D. Althoff et al. (Eds.): ECCBR 2008, LNAI 5239, pp. 18–43, 2008.

Given the objective of discovering the inherent organisation of the CBR research literature, there are three issues to be considered:

1. Should the organisation be based upon authors or papers?
2. What is the best measure of similarity to use in organising things?
3. What technique (algorithm) should be used to perform the organisation?

In large bibliometric analysis tasks, it is perhaps more conventional to use authors rather than papers as the basic unit of organisation. However, we have found that an analysis based on papers produces a clearer picture when working with a relatively small set of papers. We suggest that this is because we are partitioning a specific discipline into sub-topics, and because individual authors in the CBR area have frequently contributed to a range of different sub-topics, making an analysis based on authors more convoluted.

A variety of different measures can be used to identify relationships between papers and between authors in a collection of publications. A simple approach is to examine co-authorship relations between authors. However, in the CBR literature this approach appears to tell us more about geography than research themes. Citation links between papers are another important source of information, as they allows us to construct a network of scientific communication [3]. A related source of information, paper and author *co-citations*, has been frequently shown in bibliometric research to uncover more significant relationships than those identified using raw citation counts [4]. Text similarity, based on a "bag-of-words" representation of a corpus of papers, is yet another useful measure of similarity between research papers.

Among these different measures, we have found paper-paper co-citations to be particularly informative in the task of analysing the network formed from the publications of the CBR conference series (see Section 4). Taking co-citation as a useful means of assessing connectedness amongst research papers, it is interesting to look at the *eigenvector centrality* of overall network of papers covered in this study. The top ranked list of papers based on this criterion is presented in Section 4.1. It is interesting to compare this ranking with the list of papers as ordered by raw citation frequencies – this list is also presented in that section.

One of the objectives of this work was to checkpoint the progress of case-based reasoning research, after these last fifteen years of European and International conferences. We were particularly interested in understanding the thematic relationship between "modern case-based reasoning" and the more traditional view of case-based reasoning that dominated research prior to the commencement of the ECCBR/ICCBR series. To what extent have important new research themes emerged in the last fifteen years, for example? Is there evidence to suggest that more traditional lines of enquiry have reached a natural conclusion within the research space? With this in mind our cluster-based analysis has revealed a number of interesting results.

The good health of CBR research is supported by the frequent emergence of novel research ideas that have a history of developing into significant themes in their own right. As we explore the research groupings that have emerged in our analysis (see Section 4.2), we will highlight examples of important research

themes that have developed and matured over the past fifteen years. For example, since the early work of [5], *conversational case-based reasoning* has emerged as an important area of research that continues to attract a significant contribution at modern CBR conferences. And more recently we have seen new work in the area of *explanation* in CBR, focusing on the role that cases play when it comes to justifying decisions or recommendations to end-users; see for example [6,7,8]. Although the earliest paper in this theme is the paper by Aamodt from 1993 [9] this is still a new area of activity that has captured the attention of CBR researchers and is likely to grow in maturity over the coming years.

Of course, research themes naturally come and go with some research activities maturing to merge with the mainstream of CBR, while others appear to be more short-lived as their activity levels are seen to decline. Perhaps one of the most significant themes that has emerged in recent times has centred on the idea of *case base maintenance* – the need to actively maintain the quality of live case bases – and developed from the early work of [5,10,11,12,13]. This is a good example of a research area whose activity has now begun to reduce as maintenance techniques become well established within CBR systems; indeed this line of research has had a lasting influence on the CBR process model with a maintenance component now seen as a standard part of the CBR process [14]. More recent research in the area *diversity* — challenging the traditional similarity assumption in CBR and arguing the need for diversity among retrieved cases — seems to be heading in a similar direction: a critical mass of research from 2001 - 2004 (e.g., [15,16,17,18]) looks to be reaching a natural conclusion as the basic trade-off between similarity and diversity comes to be accepted by practitioners.

This paper begins in Section 2, with a description of the data that has been gathered for this work. The cluster analysis technique used in our work is described in Section 3. A discussion of the findings of our analysis task is presented in Section 4, and the paper finishes with some conclusions in Section 5.

2 Data Representation

Since the conception of the CBR conference series (ECCBR/ICCBR/EWCBR) in 1993, a total of 672 papers have been published by 828 individual authors. Data on these papers was gathered from the Springer online bibliographies[1] for each of the annual conference proceedings. These bibliographies are available in the form of RIS files, a tagged file format for expressing citation information, including details such as the issue title, paper titles, author lists, and abstracts for each publication in the conference series.

To determine the connections within the network of CBR publications, we submitted queries to Google Scholar[2] to retrieve the list of papers referencing each of the 672 "seed" papers. Each list contains all of the Google verified citations that a given paper had received at query submission time (December

[1] Downloaded from `http://www.springer.com`

[2] See `http://scholar.google.com`

2007). In total 7078 relevant citation links were recorded. Note that, while citation information from the supplementary (*i.e.* non-seed) set of papers was used to provide additional information regarding co-citations, only the 672 seed papers and their associated authors were considered as data objects in our analysis.

In addition to the information provided by citation links, the availability of paper titles and abstracts in the RIS format allowed us to construct an alternative view of the seed papers, in the form of a "bag-of-words" text representation. After applying standard stemming, stop-word removal and TF-IDF pre-processing techniques, the 672 conference papers were represented by feature vectors corresponding to 1487 unique terms. Similarity values between pairs of papers were computed by finding the cosine of the angle between their respective term vectors.

2.1 Co-citation Analysis

The most fundamental representation used to model scientific literature in bibliometrics is the unweighted directed citation graph, where an edge exists between the paper P_i and the paper P_j if P_i cites P_j. This graph can be represented by its asymmetric adjacency matrix $\mathbf{A}$. However, it has been established in bibliometrics research that co-citation information can be more effective in revealing the true associations between papers than citations alone [4].

The concept of co-citation analysis is illustrated in Figure 1. A direct analysis of citation shows for instance that P_1 is related to P_2. However, the fact that P_3 and P_4 are both cited by P_1 and P_2 indicates a strong relationship between these papers. In this simple example co-citation analysis suggests a weaker relationship between P_3 and P_5 and P_4 and P_5 based on co-citation in P_2. Thus co-citation has the potential to reveal indirect associations that are not explicit in the citation graph.

Consequently, a network of publications is often represented by its weighted undirected co-citation graph. This graph has a symmetric adjacency matrix defined by $\mathbf{C} = \mathbf{A}^{\mathsf{T}}\mathbf{A}$, where the off-diagonal entry C_{ij} indicates the number of papers jointly citing both P_i and P_j. Note that the entry C_{ii} on the main diagonal correspond to the total number of citations for the paper P_i.

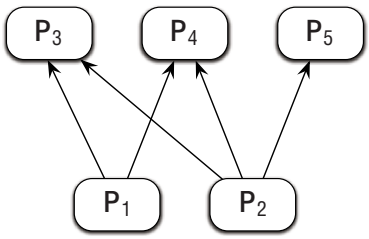

Fig. 1. Co-citation information can be more effective in revealing relationships between papers that direct citations. In this example, the fact that papers P_3 and P_4 are both cited by papers P_1 and P_2 is indicative of a relationship between them. (Note that an arrow from P_i to P_j indicates that paper P_i cites paper P_j.)

Rather than using raw co-citation values in **C** as a basis for measuring the similarity between papers, a variety of normalisation strategies have been proposed in the area of bibliometrics [19]. The *CoCit-Score*, proposed by Gmür [3], has been shown to be a particularly effective choice for clustering co-citation data. This measure allows us to compute a pairwise similarity matrix **S**, such that the similarity between a pair of papers (P_i, P_j) is given by normalising their co-citation frequency with respect to the minimum and mean of the pair's respective citation counts:

$$S_{ij} = \frac{{C_{ij}}^2}{min(C_{ii}, C_{jj}) \times mean(C_{ii}, C_{jj})} \quad (1)$$

Each entry S_{ij} is in the range $[0, 1]$, where a larger value is indicative of a stronger association between a pair of papers.

3 Cluster Analysis Techniques

A natural approach to identifying the thematic subgroups in a bibliographic network, such as the CBR conference series dataset, is to apply cluster analysis techniques. Traditional methods such as hierarchical agglomerative clustering have previously been used for this task [19]. However, a distinct drawback of these methods lies in the fact that each paper can only reside in a single branch of the tree at a given level, and can only belong to a single leaf node.

As an alternative, matrix decomposition techniques such as Non-negative Matrix Factorization (NMF) have been recently employed in the analysis of data where overlapping structures may exist [2]. Unlike other hierarchical or partitional clustering algorithms that produce disjoint (*i.e.* non-overlapping) clusters, an NMF factorisation allows each data object to potentially belong to multiple clusters to different degrees, supporting the identification of overlapping subgroups. However, there are a number of drawbacks apparent when applying NMF in practical applications, notably its sensitivity to the choice of parameter k, and the difficulty in interpreting the factors produced by the decomposition procedure.

3.1 Soft Hierarchical Clustering

We would ideally like to combine both the ability of NMF techniques to accurately identify overlapping structures, with the interpretability and visualization benefits of hierarchical techniques. Towards this end, we make use of the *Ensemble NMF* algorithm [20], which was previously applied to large protein interaction networks to address the issue of proteins belonging to more that one functional group. In the context of the CBR bibliographic network, we apply it to identify overlapping subgroups corresponding to specific areas of research within the CBR community, and to investigate how these areas have developed over the course of the conference series. The Ensemble NMF algorithm is motivated by existing unsupervised ensemble methods that have been proposed to improve

the accuracy and robustness of cluster analysis procedures by aggregating a diverse collection of different clusterings [21]. However, rather than combining hard clusterings (*i.e.* sets of disjoint, non-overlapping clusters), the algorithm involves aggregating multiple NMF factorisations. We refer to the output of this procedure as a *soft hierarchical clustering* of the data, as data objects (*e.g.* research papers) are organised into a binary tree such that they can be associated with multiple nodes in the tree to different degrees. A complete description of the Ensemble NMF algorithm is provided in Appendix A.

3.2 Assessing Paper Importance

When seeking to identify groups of related papers, the use of Ensemble NMF in conjunction with a similarity matrix constructed using a co-citation similarity function (such as Eqn. 1) is appropriate. However, the values in the resulting membership vectors will measure the level of *association* between each paper and a given cluster, rather than indicating the *importance* of the paper within that cluster. For instance, a paper may receive a high membership weight for a cluster as it is strongly related to the specific theme represented by the cluster, when in fact it has received relatively few citations in the literature.

To produce a meaningful ranking of the importance of the papers occurring in each cluster, we apply a re-weighting scheme based on the concept of centrality. In graph theory, the *degree* of a vertex in a graph refers to the number of edges incident to that vertex. A related measure, *degree centrality*, is commonly used as a means of assessing importance in social network analysis [22]. The rationale behind this measure is that the greater the degree of a vertex, the more potential influence it will exert in a network. For a weighted graph, we can compute a centrality score for a given vertex based on the sum of the edge weights on the edges incident to that vertex. For the co-citation graph with adjacency matrix $\mathbf{C}$, this will represent the sum of the number of co-citations for each paper.

Since our focus was on the identification of influential papers within each subgroup, we consider a measure of *local degree centrality* based on co-citation counts. Firstly, for each cluster node in the soft hierarchy, we assign papers to the cluster if their previous membership weight for that cluster exceeds a given threshold. We found experimentally that a threshold of 0.1 proved suitable in this context. Subsequently, for each paper deemed to belong to a given cluster, we calculate the number of co-citations between the paper and all other papers deemed to be in that cluster. To make scores comparable across different clusters, these values can be normalised with respect to the total number of unique pairs of articles in a given cluster. This yields new membership weights in the range $[0, 1]$, where a higher score indicates that a paper is more influential in the area of research covered by a specific cluster.

3.3 Back-Fitting Recent Papers

One drawback of citation analysis is that we must wait for a sufficient amount of time to pass for citations to accrue in order to identify the associations

between a paper and previously published work. As a result, most recent papers in the CBR conferences series (from 2005 onwards) did not feature strongly in the clusters generated on co-citation data. To address this issue, we propose a simple approach to "back-fit" these papers to the clusters generated with Ensemble NMF. Using the disjoint cluster memberships derived in Section 3.2, we associate each unassigned recent paper to a cluster if that paper cites three or more of the papers within the cluster. This stringent threshold led to relatively few assignments, which is desirable as we only wished to identify new papers that were strongly related to the groups discovered during the clustering process.

3.4 Labelling Clusters

The text representation described in Section 2 proved valuable as a means of summarising the content of the clusters in the soft hierarchy prior to human inspection. Cluster keywords were automatically identified by ranking the terms for each cluster based on their Information Gain [23]. 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; 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 Analysis

In this section, we discuss the analysis of the CBR dataset described in Section 2 based on the application of the Ensemble NMF algorithm. As noted previously, a variety of different measures can be used to identify groupings in a collection of publications. In our initial experiments, we applied the algorithm to four different representations of the CBR network: the raw author-author co-citation matrix, the raw paper-paper co-citation matrix, the paper-paper CoCit-score matrix, and the Cosine similarity matrix constructed from the text data. Note that co-citation links from the supplementary papers retrieved from Google Scholar (as described in Section 2) was used in the construction of the co-citation matrices.

For each data representation, 1000 ensemble members were generated using symmetric NMF, with a range $k \in [15, 20]$ used for the number of basis vectors in each factorisation. This range was chosen by inspecting the gaps between the ordered set of eigenvalues in the eigenvalue decomposition of the individual similarity matrices, as frequently applied in spectral analysis [24]. These evaluations showed that clusterings generated on the CoCit-score matrix yielded clusters that were far more informative in terms of producing meaningful thematic groupings, without containing an undue bias toward the geographical co-location of authors. Consequently, in the remainder of this paper we focus on the output of the Ensemble NMF algorithm on this particular representation.

To examine these results in detail, we developed the "NMF Tree Browser" tool, a cross-platform Java application for visually inspecting a soft hierarchy

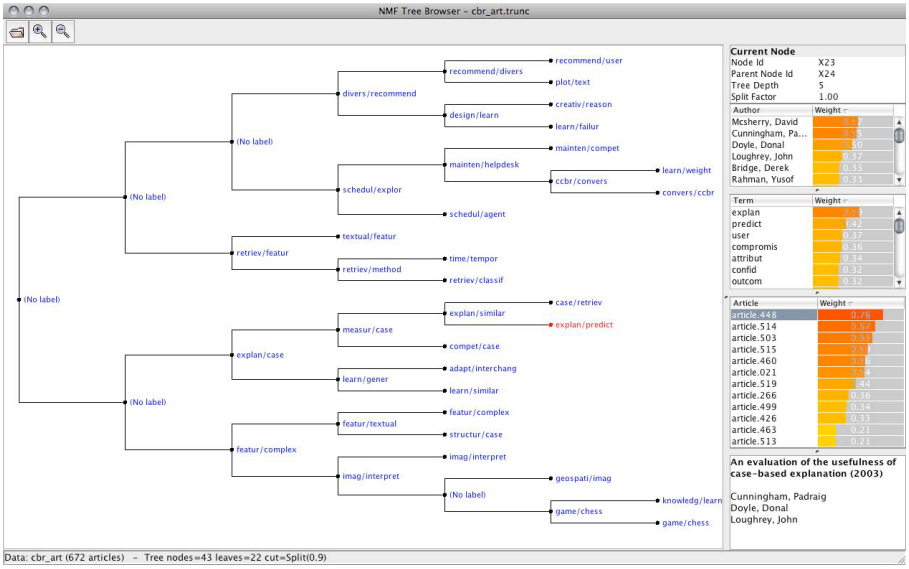

Fig. 2. Screenshot of the *NMF Tree Browser* application displaying the output of the Ensemble NMF procedure when applied to the CBR network dataset

as produced by the Ensemble NMF algorithm. The clustering is graphically arranged in a tree view, where the user can click on any node to reveal its contents, in terms of relevant papers, authors and descriptive terms. A screenshot of the main window of the application is shown in Figure 2. The application is freely available online[3], together with the data files used in our experiments.

4.1 Global Picture

In this section we look at the salient global statistics for the complete set of papers presented at the conference series since 1993. Some statistics on citations are provided in Table 1. It is interesting to note that ICCBR papers are no more significant (in terms of citations) than ECCBR papers. In fact the mean and median number of citations per paper is marginally higher for ECCBR than for ICCBR. We feel this validates our strategy of treating these as a homogenous set of papers.

Given that the main findings in this paper entail a clustering of the papers based on co-citation links, it is interesting to see which papers are most 'central' to the overall collection based on these co-citation links. Following the literature on centrality in social network analysis, we selected eigenvector centrality and degree centrality as appropriate measures for this exercise [22]. Table 2 shows the top 10 papers ranked by eigenvector centrality. This table also shows a count of co-citations for these papers – this corresponds to degree centrality and correlates

[3] The browser tool and data files can be downloaded from `http://mlg.ucd.ie/cbr`

Table 1. A comparison of overall citation statistics between ECCBR and ICCBR

Conference	No. Papers	Maximum	Mean	Median
ECCBR	305	92	11.01	6
ICCBR	367	137	10.14	5

well with eigenvector centrality. A further ranked list with papers ranked by raw citation count is shown in Table 3. The evidence from these tables is that the most important paper in the collection is "Weighting Features" by Wettschereck & Aha [25]. These two lists of prominent papers are useful in that they do appear to encapsulate the main themes in CBR research over the last 15 years.

An obvious shortcoming of the analysis reported here is that it is restricted to papers presented at the international conferences since 1993 only. This excludes a number of important publications that have greatly influenced the field and are strongly linked to the papers that have been covered. Perhaps the most prominent example of this is the paper by Aamodt & Plaza [14] – this is the definitive citation for the CBR cycle which shapes the way we think about the CBR process. Another important influence on CBR research has been Richter's "knowledge containers" idea that he introduced in an invited talk at ICCBR'95. Unfortunately this work is not included in the CBR conference proceedings, but is described elsewhere [26].

4.2 Analysis of Subgroups

As a result of this analysis we have been able to identify a number of important research themes within the CBR literature, corresponding to cohesive clusters in the soft hierarchy produced by Ensemble NMF. We refer to these as the *modern* CBR themes, since they reflect how research focus has shifted over the past fifteen years, and they clearly differ from more traditional CBR themes such as representation and indexing, retrieval and similarity, adaptation, learning, analogy, planning and design etc. In this section we briefly review and discuss these *modern* CBR themes.

In addition, Figures 3 and 4 provide timelines which profile each theme in terms of its core papers, and their relative centrality and impact for the duration of the conference series. Each timeline shows the papers in a selected cluster (*i.e.* modern research theme) in three dimensions: the year of the conference (x-axis), the centrality of the paper in the cluster (y-axis), and the number of citations for that paper (depicted by the size of the disc representing the paper). For reasons of scale, papers with more than 50 citations are represented by a disc of size 50 – this only happens for 3% of papers. Since eigenvector centrality can be unreliable for small clusters, paper importance is measured by [0-1]-normalised local degree centrality, as previously defined in Section 3.2. It can be seen from the figures that different clusters have different importance profiles. This can be interpreted to mean than clusters such as Case-Base Maintenance are more compact and cohesive than clusters such as Case Retrieval. The timelines also show papers that have been back-fitted to the clusters as described in Section 3.3.

Table 2. A ranked list of the top 10 papers in the overall collection based on eigenvector centrality. The total number of citations and the number of co-citations for these papers is also shown.

#	Paper	Year	Citations	Co-cites
1	*Weighting features* Wettschereck & Aha	1995	137	522
2	*Modelling the competence of case-bases* Smyth & McKenna	1998	92	525
3	*Refining conversational case libraries* Aha & Breslow	1997	117	518
4	*Maintaining unstructured case bases* Racine & Yang	1997	72	469
5	*Using introspective learning to improve retrieval in CBR: A case study in air traffic control* Bonzano *et al.*	1997	74	473
6	*Similarity vs. diversity* Smyth & McClave	2001	72	452
7	*Building compact competent case-bases* Smyth & McKenna	1999	64	399
8	*Categorizing case-base maintenance: dimensions and directions* Leake & Wilson	1998	82	322
9	*Diversity-conscious retrieval* McSherry	2002	44	362
10	*Similarity measures for object-oriented case representations* Bergmann & Stahl	1998	66	403

These papers are represented by blue discs. Note that all papers mentioned in this section are labelled with their corresponding reference number.

Recommender Systems and Diversity: Recent research interest in recommender systems has provided the impetus for a new take on one of the long-held assumptions that has underpinned case-based reasoners, namely the *similarity assumption*. The similarity assumption states that the similarity between the target specification (query) and cases in the case base is the primary retrieval constraint in CBR systems; in other words, that cases should be selected and ranked for retrieval in terms of their similarity to the target specification. The idea that this assumption does not always hold is an important theme within the area of recommender systems (both single-shot and conversational). The work of [15] argued that an exclusive focus on similar cases can lead to the retrieval of a homogeneous set of case that fail to offer the user a diverse set of alternatives, which is often an important consideration in many recommendation scenarios. In addition [15] first introduced the notion of a diversity conscious approach to case retrieval, with a view to producing more diverse retrieval-sets that provide a better set of alternatives to a user. This work captured the interest on a

Table 3. A ranked list of the top 10 papers in the overall collection based on total citation count

#	Paper	Year	Citations
1	*Weighting features* Aha & Wettschereck	1995	137
2	*Refining conversational case libraries* Aha & Breslow	1997	117
3	*Modelling the competence of case-bases* Smyth & McKenna	1998	92
4	*Categorizing case-base maintenance: dimensions and directions* Leake & Wilson	1998	82
5	*Using k-d trees to improve the retrieval step in case-based reasoning* Althoff *et al.*	1993	76
6	*Using introspective learning to improve retrieval in CBR: a case study in air traffic control* Bonzano *et al.*	1997	74
7	*Explanation-driven case-based reasoning* Aamodt	1993	72
8	*Maintaining unstructured case bases* Racine & Yang	1997	72
9	*Similarity vs. diversity* Smyth & McClave	2001	72
10	*Cases as terms: A feature term approach to the structured representation of cases* Plaza	1995	70

number of CBR researchers with the work of [17,18,27,28] providing a number of extensions to this original diversity work.

This particular theme is notable because of the relatively large number of highly cited, very central papers over a short and recent time period as shown in Figure 3. The first two papers in this cluster [29,30] are early papers on recommender systems that are also prominent in the Conversional CBR cluster described later. This shows the benefit of a clustering strategy that allows objects to belong to more than one cluster.

Case-Base Maintenance: One cluster of research that stands out particularly well in our co-citation analysis concerns the area of *case base maintenance*. In fact, this line of research has had a lasting impact on the landscape of case-based reasoning, with maintenance now viewed an a standard component of modern CBR systems. At the heart of case-base maintenance is the idea that the quality of the case base as a whole needs to be actively managed, to ensure that erroneous cases can be identified, if not removed, and so that redundancy may be reduced as a way to stave of the impact of the utility problem. A key publication in this area of research is the work of Leake & Wilson [10] which attempted, for the first time, to categorise the various factors that influence case base maintenance as well as laying out the challenges and opportunities for future research.

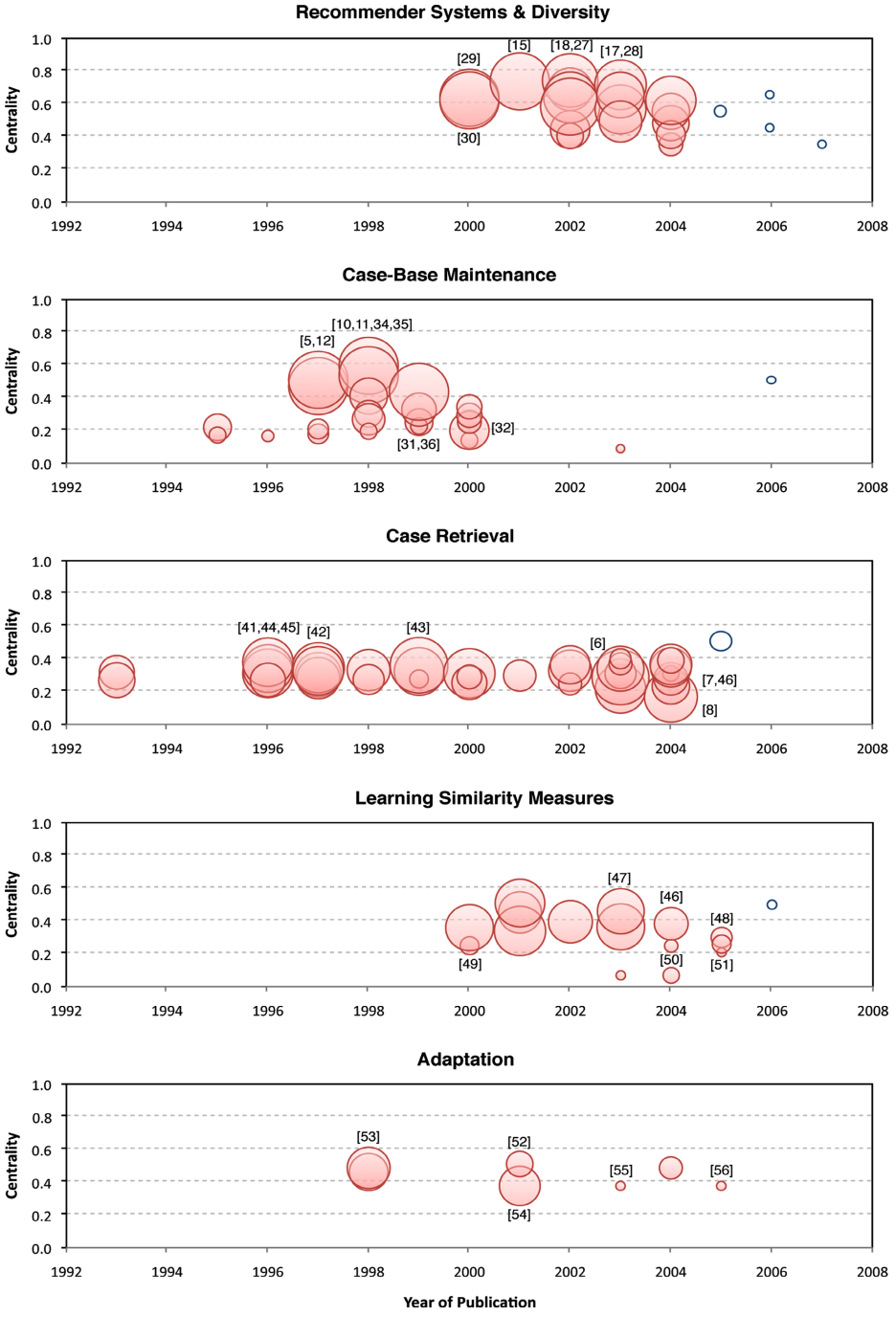

Fig. 3. Timeline plots for selected leaf node clusters. The size of the disc for each paper indicates its number of citations, and the position on the y-axis indicates its centrality.

Subsequently, many researchers have focused on developing specific maintenance techniques, some looking at different ways to measure case quality (e.g. [11,31,32]), while others propose novel techniques for pruning or editing or otherwise optimising the case base (e.g. [12,5,33,34,35,36]). It is worth noting that this research area has evolved from a number of papers that have been published outside of the ICCBR/ECCBR and, as such, are beyond the scope of this analysis. These papers include early work on understanding the *utility problem* [37] in a CBR context [38,39,40], especially the idea that traditional ML-style strategies for coping with the utility problem, namely the outright deletion of learned knowledge, might not be appropriate in a CBR setting [13]. Once again this cluster is characterised by a relatively large number of papers over a relatively short period of time. It is also interesting to note that a small number of these papers attract the lion's share of citations (see Table 2), with other works playing a much less central role by exploring different aspects of the case base maintenance. It is also notable that there has not been much new research in this area in recent years. Perhaps this is an indication that this line of research has now become common practice in CBR, with effective solutions already available.

Case Retrieval: From the beginning, case-based reasoning research has been heavily influenced by the so-called *similarity assumption* — that the best case to retrieve is that which is most similar to the target problem — and the early years of CBR research were guided by cognitively-plausable similarity assessment techniques. Contemporary CBR research has adopted a much more flexible position when it comes to case retrieval and similarity assessment. Many researchers have argued that similarity alone is rarely enough to guide retrieval, for example, while others have pointed out that cases can be retrieved for purposes other than problem solving (e.g., explanation). This body of research is evident within our analysis as a cluster that covers a broad spectrum of contributions over an extended period of time. These include early work on the foundations of case retrieval and similarity [41,42], and the proposal of novel retrieval methods that go beyond a pure similarity-based approach [43,44,45,46], to more recent work on case explanation [7,8], where the job of retrieval is not to select a case that will help to justify or explain a conclusion, a case which might not be the most similar to a given problem [6].

Learning Similarity Measures: The importance of retrieval and similarity in CBR research is evidenced by the emergence of two clusters of research that speak to these topics. Above we have discussed research related to the role of similarity in retrieval and in this section we briefly highlight the second cluster which is dominated by work on the learning of similarity measures for case retrieval. The work of Armin Stahl is particularly prominent in this cluster, with a number of important papers covering the learning of feature weights [47], the role of background knowledge in similarity learning [46], as well as a proposal for a formal framework for learning similarity measures based on a generalised model of CBR [48]. It is also worth highlighting some of the related research that appears in this cluster, which focuses on the role of user preferences in similarity with

research by [49,50,51], for example, looking at different approaches to harnessing user profiles and user preferences in similarity-based retrieval.

Adaptation: One of the smaller clusters of research activity that has emerged from our analysis is in the area of *case adaptation*. Despite a strong showing in the early years of CBR, work in the area of adaptation is now far less prominent, at least within the ECCBR/ICCBR series. And while the level of activity on this topic has promoted some to proclaim the death of case adaptation there are clear signs that researchers have not given up on this most challenging of CBR tasks. This cluster, for example, reflects recent work in the area of case adaptation and includes practical work on domain specific adaptation techniques [52] and more general approaches to case adaptation such as the work of [53,54,55,56]

Image Analysis: CBR researchers will not be surprised to see that image analysis (particularly medical image analysis) has been identified as a distinct research theme in the CBR literature. The earliest paper in the cluster that has been identified is from ICCBR'95 by Macura & Macura [57], which describes the application of CBR in the area of radiology imaging. Two other central papers in this cluster are the paper on using CBR for image segmentation by Perner [58] and a paper describing a CBR architecture for medical image understanding by Grimes & Aamodt [59]. This cluster also includes two papers on geospatial image analysis, although the dominant theme in this area of CBR has been medical image analysis. Given that significant research challenges still exist in image analysis it is interesting that the clustering has attached few very recent papers to this theme. The process of back-fitting recent papers as described in Section 3.3 has added only one paper. Part of the explanation for this is that some of the research activity in this area is reported outside the CBR conferences. Surely this is an area of research that warrants more attention from the CBR community.

Textual CBR: Ensemble NMF co-citation clustering identifies a theme that is characterised by terms such as *textual, CCBR, text, question* and *taxonomy*. An examination of the papers in this cluster shows that it covers Textual CBR. While the earliest paper in this theme is from Brüninghaus & Ashley in 1997 [60] most of the material is from recent years. So this is a new but still well established theme in CBR research. Some key papers in this cluster are the 2002 paper by Gupta *et al.* [61] and the 2004 paper by Wiratunga *et al.* [62]. It is interesting that if the clustering is allowed to further divide the corpus then this cluster splits into two distinct sub-groups: one pertaining to textual CBR [60,63,64,62], and another pertaining to conversational CBR [61,65].

Conversational CBR: The cluster analysis reveals some interesting insights into research on conversational CBR (CCBR). In fact CCBR papers are divided into two sub-groups: one is associated with textual CBR in the overall cluster hierarchy and the other is linked to learning and induction. The key papers in the *textual* side of conversational CBR have been described already in the previous section. Some central papers from the *learning* side of CCBR are the

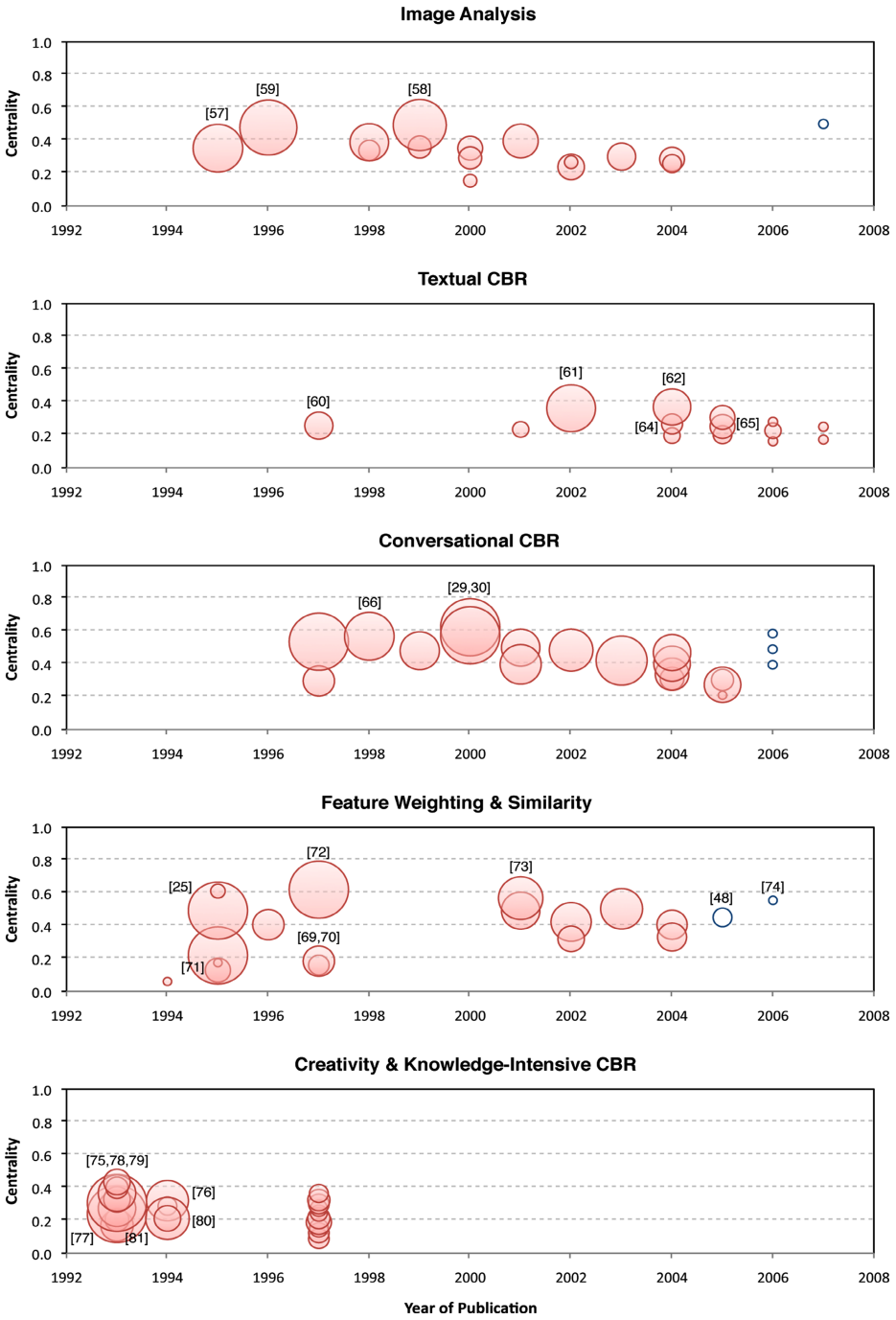

Fig. 4. Timeline plots for selected leaf node clusters (continued)

2000 paper by Doyle & Cunningham [29], the 2000 paper by Göker & Thompson [30] and the 1998 paper by Aha *et al.* [66]. This is a significant cluster that contains seventeen papers, most of which have attracted an impressive number of citations. In addition to this link to conversational CBR, this cluster also links to the Recommender Systems and Diversity theme where the papers [29,30] are also prominent. The back-fitting process has attached another three papers to this cluster. It is clear from the timelines in Figure 4 that this research area is in rude good health with considerable activity in the area.

Feature Weighting and Similarity: In fact, the clustering further divides this cluster into two sub-groups, one on fault diagnosis and another on feature weight learning. The former is unusual in that it contains no recent papers; there is one paper from 2000 [67], and before that the most recent papers are from 1997 [68,69,70]. There have continued to be papers on diagnosis in the research literature but it does not seem to connect with this literature through co-citation. Instead the clustering process has connected more recent papers on diagnosis with research on textual CBR or with work on similarity for structured representations. Two representative papers that describe the work on fault diagnosis in this cluster are the work of Netten & Vingerhoeds [71], and Jarmulak *et al.* [69].

The other part of this cluster comprises papers on feature weight learning. The seminal paper in this collection is the paper by Wettschereck & Aha from 1995 on "Weighting Features" [25]. This is also the most central and significant paper is the whole 15 year collection (see Section 4.1). Other central papers in this cluster are the paper on using introspective learning to learn feature weights by Bonzano *et al.* [72] and the work by Stahl on learning feature weights from case order feedback [73]. While activity in this area may be slowing down, there appears to be ongoing work as the process of back-fitting papers has linked two papers from 2004 and 2005 to this cluster [48,74].

Creativity & Knowledge-Intensive CBR: One of the more remarkable groupings revealed by the clustering process is the one we have called "Creativity & Knowledge Intensive CBR". The keywords associated with this cluster are *creative, reason, design, rule, interpolation, tune, represent, model, integrate* and *adapt* and the influential papers include [75,76,77,78,79,80]. This cluster is unusual in that the most recent papers are from 1997. Thus, it represents a body of research that has either waned or been taken up in other areas. An analysis of the prominent papers in this cluster supports the impression created by the list of terms above that this cluster covers research on knowledge intensive CBR and links with earlier work on analogy and model-based reasoning. This cluster includes papers on CBR as a creative problem solving process; the first paper in this sequence is the invited paper from the 1993 conference by Kolodner on "Understanding Creativity: A Case-Based Approach" [81].

It would be wrong to think of this as a strand of CBR research that did not 'work out'. Rather, some of the papers in this cluster have proved influential in other areas within CBR. For instance, the paper by Bunke and Messmer, on "Similarity Measures for Structured Representations" [77] is a very influential

paper in work on similarity and is still cited today. Furthermore the connection between CBR and induction that went on to be a major theme in CBR in the late 1990s is a prominent theme in some of the papers in this cluster [76,82]. The paper by Smyth and Keane on retrieving adaptable cases [78] marked the beginning of a strand of research that offered a new perspective on case retrieval. On the other hand, the view of CBR as a model of creativity does seem to have waned. Perhaps this is no surprise as, to a large extent, the modern view of CBR is one the emphasises retrieval rather than adaptation and, arguably, creativity demands a significant measure of adaptation by definition. The early work of creativity [81] stems from a time when there was a more optimistic view of the potential for automated adaptation, and the lack of significant research activity in the area of adaptation (as discussed above) suggests that this view is no longer held.

5 Conclusion

In this work, we have set out to review the last fifteen years of CBR research with a view to understanding how major research themes have developed and evolved over this extended period of time. Unlike many more traditional research reviews, which tend to adopt a top-down style of analysis based on long-accepted thematic norms, we have instead opted for a bottom-up style of analysis. Our intuition has always been that CBR research tends to be dynamic, with new research themes emerging on a reasonably regular basis, and as such a pre-canned top-down analysis would run the risk of missing important developments that fall outside of the traditional themes.

Our bottom-up analysis has focused on mining the relationships between papers and authors from the fifteen years of international CBR conferences. The results confirm that modern CBR research is characterised by a set of research themes that are significantly different from those that would have characterised the early years of the field. We have identified strong clusters of activity in areas such as Recommender Systems & Diversity, Textual CBR, Case-Base Maintenance and Conversational CBR, which we believe to be characteristic of modern CBR research. Interestingly, many of the more traditional research themes do not feature prominantly in the clusters of research that have emerged from our analysis. For example, the traditional themes of *representation and indexing*, *analogy*, *architectures*, and *design and planning* are conspicuous by their absence and even critical areas of research such as *adaptation* or *similarity and retrieval* have either become less active or have fundamentally changed their emphasis.

It is also pleasing to note from Figures 3 and 4 that new themes can emerge (e.g. Recommender Systems & Diversity), and that research activity in an area can come to a close (e.g. Case-Base Maintenance), as it matures to deliver effective solutions to the community. This can be considered a sign of a healthy research area.

The choice of a clustering algorithm that produces a "soft" hierarchical organisation, allowing the identification of localised groupings where papers may

belong to more than one cluster, has proved effective. This has revealed some interesting links and overlaps between areas. For instance, overlaps between the areas of Textual CBR and Conversational CBR, and between Conversational CBR and Recommender Systems. It has also revealed the two aspects of Conversational CBR, the textual side and the learning side.

In this paper we have limited our discussion to the ten most prominent research themes, largely based on the size of the cluster (in terms of papers published). It is worth highlighting that a number of more minor clusters have also been identified, including:

- CBR on temporal problems: *time, temporal, prediction, series.*
- Games and chess: *game, chess, automatic, sequential.*
- Scheduling and agents: *schedule, agent, exploration.*
- Structural cases: *structural, case, induction, logic.*

Clearly these clusters also represent important and interesting lines of research. Work in the area of games and chess, while something of a *niche* area, has been part of CBR research since 1995 [83], and continues to attract research interest. Others clusters such as *CBR on temporal problems* cover a broad spectrum of work dealing with a range of issues, such as using CBR to predict time-series [84] and the representation of temporal knowledge in case-based prediction [85] to more recent work on so-called historical case-based reasoning [86]. There is no doubt that these themes are worthy of additional research, and a further exploration of the papers in these clusters will no doubt lead to further fruitful insights into the ever-changing landscape of CBR research.

References

1. Greene, D., Cunningham, P., Mayer, R.: Unsupervised learning and clustering. In: Cord, M., Cunningham, P. (eds.) Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval, pp. 51–90. Springer, Heidelberg (2008)
2. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
3. Gmür, M.: Co-citation analysis and the search for invisible colleges: A methodological evaluation. Scientometrics 57, 27–57 (2003)
4. White, H., Griffith, C.: Author Cocitation: A Literature Measure of Intellectual Structure. J. ASIS 32, 163–171 (1981)
5. Aha, D., Breslow, L.: Refining conversational case libraries. Case-Based Reasoning Research and Development, 267–278 (1997)
6. Cunningham, P., Doyle, D., Loughrey, J.: An evaluation of the usefulness of case-based explanation. Case-Based Reasoning Research and Development, 1065 (2003)
7. McSherry, D.: Explaining the pros and cons of conclusions in cbr. Advances in Case-Based Reasoning, 317–330 (2004)
8. Doyle, D., Cunningham, P., Bridge, D., Rahman, Y.: Explanation oriented retrieval. Advances in Case-Based Reasoning, 157–168 (2004)
9. Aamodt, A.: Explanation-driven case-based reasoning. Advances in Case-Based Reasoning (1993)

10. Leake, D.B., Wilson, D.C.: Categorizing case-base maintenance: Dimensions and directions. Advances in Case-Based Reasoning, 196 (1998)
11. Smyth, B., McKenna, E.: Modelling the competence of case-bases. Advances in Case-Based Reasoning, 208 (1998)
12. Racine, K., Yang, Q.: Maintaining unstructured case bases. Case-Based Reasoning Research and Development, 553–564 (1997)
13. Smyth, B., Keane, M.T.: Remembering to forget: A competence-preserving case deletion policy for case-based reasoning systems. In: IJCAI, pp. 377–383 (1995)
14. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 7, 39–59 (1994)
15. Smyth, B., McClave, P.: Similarity vs. diversity. In: Aha, D.W., Watson, I. (eds.) ICCBR 2001. LNCS (LNAI), vol. 2080, p. 347. Springer, Heidelberg (2001)
16. McSherry, D.: Diversity-conscious retrieval. In: Craw, S., Preece, A.D. (eds.) EC-CBR 2002. LNCS (LNAI), vol. 2416, pp. 27–53. Springer, Heidelberg (2002)
17. McGinty, L., Smyth, B.: On the role of diversity in conversational recommender systems. Case-Based Reasoning Research and Development, 1065 (2003)
18. Bridge, D., Ferguson, A.: Diverse product recommendations using an expressive language for case retrieval. In: Craw, S., Preece, A.D. (eds.) ECCBR 2002. LNCS (LNAI), vol. 2416, pp. 291–298. Springer, Heidelberg (2002)
19. He, Y., Cheung Hui, S.: Mining a Web Citation Database for author co-citation analysis. Information Processing and Management 38, 491–508 (2002)
20. Greene, D., Cagney, G., Krogan, N., Cunningham, P.: Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-Protein Interactions. Bioinformatics (2008)
21. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining partitionings. In: Proc. Conference on Artificial Intelligence (AAAI 2002), pp. 93–98. AAAI/MIT Press (2002)
22. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
23. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Fisher, D.H. (ed.) Proc. 14th International Conference on Machine Learning (ICML 1997), Nashville, US, pp. 412–420. Morgan Kaufmann Publishers, San Francisco (1997)
24. Ng, A., Jordan, M., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm. Advances in Neural Information Processing 14, 849–856 (2001)
25. Wettschereck, D., Aha, D.: Weighting features. Case-Based Reasoning Research and Development, 347–358 (1995)
26. Richter, M.M.: Introduction. In: Lenz, M., Bartsch-Spörl, B., Burkhard, H.-D., Wess, S. (eds.) Case-Based Reasoning Technology. LNCS (LNAI), vol. 1400, pp. 1–16. Springer, Heidelberg (1998)
27. Mougouie, B., Richter, M.M., Bergmann, R.: Diversity-conscious retrieval from generalized cases: A branch and bound algorithm. Case-Based Reasoning Research and Development, 1064 (2003)
28. McSherry, D.: Similarity and compromise. Case-Based Reasoning Research and Development, 1067 (2003)
29. Doyle, M., Cunningham, P.: A dynamic approach to reducing dialog in on-line decision guides. In: Blanzieri, E., Portinale, L. (eds.) EWCBR 2000. LNCS (LNAI), vol. 1898, pp. 323–350. Springer, Heidelberg (2000)
30. Goker, M., Thompson, C.: Personalized conversational case-based recommendation. In: Blanzieri, E., Portinale, L. (eds.) EWCBR 2000. LNCS (LNAI), vol. 1898, pp. 29–82. Springer, Heidelberg (2000)

31. Portinale, L., Torasso, P., Tavano, P.: Speed-up, quality and competence in multi-modal case-based reasoning. In: Althoff, K.-D., Bergmann, R., Branting, L.K. (eds.) ICCBR 1999. LNCS (LNAI), vol. 1650, p. 718. Springer, Heidelberg (1999)
32. Reinartz, T., Iglezakis, I., Roth-Berghofer, T.: On quality measures for case base maintenance. In: Blanzieri, E., Portinale, L. (eds.) EWCBR 2000. LNCS (LNAI), vol. 1898, pp. 247–259. Springer, Heidelberg (2000)
33. Smyth, B.: Competence models and their applications. In: Blanzieri, E., Portinale, L. (eds.) EWCBR 2000. LNCS (LNAI), vol. 1898, pp. 1–2. Springer, Heidelberg (2000)
34. Heister, F., Wilke, W.: An architecture for maintaining case-based reasoning systems. Advances in Case-Based Reasoning, 221 (1998)
35. Surma, J., Tyburcy, J.: A study on competence-preserving case replacing strategies in case-based reasoning. Advances in Case-Based Reasoning, 233 (1998)
36. Munoz-Avila, H.: A case retention policy based on detrimental retrieval. In: Althoff, K.-D., Bergmann, R., Branting, L.K. (eds.) ICCBR 1999. LNCS (LNAI), vol. 1650, p. 721. Springer, Heidelberg (1999)
37. Minton, S.: Quantitative results concerning the utility of explanation-based learning. Artif. Intell. 42, 363–391 (1990)
38. Ram Jr., A., Francis, A.G.: The utility problem in case-based reasoning. In: Proceedings AAAI 1993 Case-Based Reasoning Workshop (1993)
39. Smyth, B., Cunningham, P.: The utility problem analysed. Advances in Case-Based Reasoning, 392–399 (1996)
40. Ram Jr., A., Francis, A.G.: A comparitive utility analysis of case-based reasoning and control-rule learning systems. In: Lavrač, N., Wrobel, S. (eds.) ECML 1995. LNCS, vol. 912, pp. 138–150. Springer, Heidelberg (1995)
41. Osborne, H., Bridge, D.: A case base similarity framework. Advances in Case-Based Reasoning, 309–323 (1996)
42. Osborne, H., Bridge, D.: Similarity metrics: A formal unification of cardinal and non-cardinal similarity measures. Case-Based Reasoning Research and Development, 235–244 (1997)
43. Smyth, B., McKenna, E.: Footprint-based retrieval. In: Althoff, K.-D., Bergmann, R., Branting, L.K. (eds.) ICCBR 1999. LNCS (LNAI), vol. 1650, p. 719. Springer, Heidelberg (1999)
44. Schaaf, J.: Fish and shrink. a next step towards efficient case retrieval in large scaled case bases. Advances in Case-Based Reasoning, 362–376 (1996)
45. Lenz, M., Burkhard, H., Bruckner, S.: Applying case retrieval nets to diagnostic tasks in technical domains. Advances in Case-Based Reasoning, 219–233 (1996)
46. Gabel, T., Stahl, A.: Exploiting background knowledge when learning similarity measures. Advances in Case-Based Reasoning, 169–183 (2004)
47. Stahl, A., Gabel, T.: Using evolution programs to learn local similarity measures. Case-Based Reasoning Research and Development, 1064 (2003)
48. Stahl, A.: Learning similarity measures: A formal view based on a generalized cbr model. Case-Based Reasoning Research and Development, 507–521 (2005)
49. Gomes, P., Bento, C.: Learning user preferences in case-based software reuse. In: Blanzieri, E., Portinale, L. (eds.) EWCBR 2000. LNCS (LNAI), vol. 1898, pp. 112–123. Springer, Heidelberg (2000)
50. Bradley, K., Smyth, B.: An architecture for case-based personalised search. Advances in Case-Based Reasoning, 518–532 (2004)
51. Hayes, C., Avesani, P., Baldo, E., Cunningham, P.: Re-using implicit knowledge in short-term information profiles for context-sensitive tasks. Case-Based Reasoning Research and Development, 312–326 (2005)

52. Bandini, S., Manzoni, S.: Cbr adaptation for chemical formulation. In: Aha, D.W., Watson, I. (eds.) ICCBR 2001. LNCS (LNAI), vol. 2080, p. 634. Springer, Heidelberg (2001)
53. McSherry, D.: An adaptation heuristic for case-based estimation. Advances in Case-Based Reasoning, 184 (1998)
54. Neagu, N., Faltings, B.: Exploiting interchangeabilities for case adaptation. In: Aha, D.W., Watson, I. (eds.) ICCBR 2001. LNCS (LNAI), vol. 2080, p. 422. Springer, Heidelberg (2001)
55. Neagu, N., Faltings, B.: Soft interchangeability for case adaptation. Case-Based Reasoning Research and Development, 1066 (2003)
56. Tonidandel, F., Rillo, M.: Case adaptation by segment replanning for case-based planning systems. Case-Based Reasoning Research and Development, 579–594 (2005)
57. Macura, R., Macura, K.: Macrad: Radiology image resource with a case-based retrieval system. Case-Based Reasoning Research and Development, 43–54 (1995)
58. Perner, P.: An architecture for a cbr image segmentation system. In: Althoff, K.-D., Bergmann, R., Branting, L.K. (eds.) ICCBR 1999. LNCS (LNAI), vol. 1650, p. 724. Springer, Heidelberg (1999)
59. Grimnes, M., Aamodt, A.: A two layer case-based reasoning architecture for medical image understanding. Advances in Case-Based Reasoning, 164–178 (1996)
60. Bruninghaus, S., Ashley, K.D.: Using machine learning for assigning indices to textual cases. Case-Based Reasoning Research and Development, 303–314 (1997)
61. Gupta, K.M., Aha, D.W., Sandhu, N.: Exploiting taxonomic and causal relations in conversational case retrieval. In: Craw, S., Preece, A.D. (eds.) ECCBR 2002. LNCS (LNAI), vol. 2416, pp. 175–182. Springer, Heidelberg (2002)
62. Wiratunga, N., Koychev, I., Massie, S.: Feature selection and generalisation for retrieval of textual cases. Advances in Case-Based Reasoning, 806–820 (2004)
63. Bruninghaus, S., Ashley, K.D.: The role of information extraction for textual cbr. In: Aha, D.W., Watson, I. (eds.) ICCBR 2001. LNCS (LNAI), vol. 2080, p. 74. Springer, Heidelberg (2001)
64. Lamontagne, L., Lapalme, G.: Textual reuse for email response. Advances in Case-Based Reasoning, 242–256 (2004)
65. Gu, M., Aamodt, A.: A knowledge-intensive method for conversational cbr. Case-Based Reasoning Research and Development, 296–311 (2005)
66. Aha, D.W., Maney, T., Breslow, L.A.: Supporting dialogue inferencing in conversational case-based reasoning. Advances in Case-Based Reasoning, 262 (1998)
67. Vollrath, I.: Handling vague and qualitative criteria in case-based reasoning applications. In: Blanzieri, E., Portinale, L. (eds.) EWCBR 2000. LNCS (LNAI), vol. 1898, pp. 403–444. Springer, Heidelberg (2000)
68. Faltings, B.: Probabilistic indexing for case-based prediction. Case-Based Reasoning Research and Development, 611–622 (1997)
69. Jarmulak, J., Kerckhoffs, E., van't Veen, P.: Case-based reasoning in an ultrasonic rail-inspection system. In: Case-Based Reasoning Research and Development, pp. 43–52 (1997)
70. Trott, J., Leng, B.: An engineering approach for troubleshooting case bases. Case-Based Reasoning Research and Development, 178–189 (1997)
71. Netten, B., Vingerhoeds, R.: Large-scale fault diagnosis for on-board train systems. Case-Based Reasoning Research and Development, 67–76 (1995)
72. Bonzano, A., Cunningham, P., Smyth, B.: Using introspective learning to improve retrieval in cbr: A case study in air traffic control. Case-Based Reasoning Research and Development, 291–302 (1997)

73. Stahl, A.: Learning feature weights from case order feedback. In: Aha, D.W., Watson, I. (eds.) ICCBR 2001. LNCS (LNAI), vol. 2080, p. 502. Springer, Heidelberg (2001)
74. Stahl, A.: Combining case-based and similarity-based product recommendation. Advances in Case-Based Reasoning, 355–369 (2006)
75. Arcos, J.L., Plaza, E.: A reflective architecture for integrated memory-based learning and reasoning. Advances in Case-Based Reasoning (1993)
76. Armengol, E., Plaza, E.: Integrating induction in a case-based reasoner. Advances in Case-Based Reasoning, 2–17 (1994)
77. Bunke, H., Messmer, B.: Similarity measures for structured representations. Advances in Case-Based Reasoning (1993)
78. Smyth, B., Keane, M.: Retrieving adaptable cases: The role of adaptation knowledge in case retrieval. Advances in Case-Based Reasoning (1993)
79. Nakatani, Y., Israel, D.: Tuning rules by cases. Advances in Case-Based Reasoning (1993)
80. Richards, B.: Qualitative models as a basis for case indices. Advances in Case-Based Reasoning, 126–135 (1994)
81. Kolodner, J.: Understanding creativity: A case-based approach. Advances in Case-Based Reasoning (1993)
82. Sebag, M., Schoenauer, M.: A rule-based similarity measure. Advances in Case-Based Reasoning (1993)
83. Flinter, S., Keane, M.: On the automatic generation of case libraries by chunking chess games. Case-Based Reasoning Research and Development, 421–430 (1995)
84. Nakhaeizadeh, G.: Learning prediction of time series - a theoretical and empirical comparison of cbr with some other approaches. Advances in Case-Based Reasoning (1993)
85. Jære, M.D., Aamodt, A., Skalle, P.: Representing temporal knowledge for case-based prediction. In: Craw, S., Preece, A.D. (eds.) ECCBR 2002. LNCS (LNAI), vol. 2416, pp. 225–234. Springer, Heidelberg (2002)
86. Ma, J., Knight, B.: A framework for historical case-based reasoning. Case-Based Reasoning Research and Development, 1067 (2003)
87. Ding, C., He, X.: On the Equivalence of Non-negative Matrix Factorization and Spectral Clustering. In: Jonker, W., Petković, M. (eds.) SDM 2005. LNCS, vol. 3674. Springer, Heidelberg (2005)
88. Opitz, D.W., Shavlik, J.W.: Generating accurate and diverse members of a neural-network ensemble. Advances in Neural Information Processing Systems 8, 535–541 (1996)
89. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10, 1299–1319 (1998)
90. Ding, C., He, X.: Cluster merging and splitting in hierarchical clustering algorithms. In: Proc. IEEE International Conference on Data Mining (ICDM 2002), p. 139 (2002)
91. Giurcaneanu, C.D., Tabus, I.: Cluster structure inference based on clustering stability with applications to microarray data analysis. EURASIP Journal on Applied Signal Processing 1, 64–80 (2004)

Appendix A: Ensemble NMF Algorithm

This appendix describes the operation of the Ensemble NMF clustering algorithm that was used in the analysis described in Section 4. The approach is

suitable for the identification of localised structures in sparse data, represented in the form of a non-negative pairwise similarity matrix, such as the co-citation matrix of the CBR network defined by Eqn. 1. The algorithm consists of two distinct phases: a *generation phase* in which a collection of NMF factorisations is produced (*i.e.* the members of the ensemble), and an *integration phase* where these factorisations are aggregated to produce a final soft hierarchical clustering of the data.

A.1 Ensemble Generation Phase

Given a dataset consisting of n data objects (*e.g.* research papers), the generation phase of the ensemble process involves the production of a collection of τ "base" clusterings. These clusterings represent the individual members of the ensemble. Since we are interested in combining the output of multiple matrix factorisations, each member will take the form of a non-negative $n \times k_i$ matrix factor $\mathbf{V}_i$, such that k_i is the number of basis vectors (*i.e.* clusters) specified for the i-th factorisation procedure.

To generate the collection of base clusterings, we employ the symmetric NMF algorithm proposed by Ding *et al.* [87]. This algorithm decomposes a non-negative pairwise similarity matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ to produce a factor $\mathbf{V}$ by minimising the objective function given by the Frobenius norm:

$$\min_{\mathbf{V}>0} \left|\left|\mathbf{S} - \mathbf{V}\mathbf{V}^{\mathsf{T}}\right|\right|_{\mathsf{F}}^{2} \tag{2}$$

The optimal factor can be approximated by starting with an initial randomly-generated factor and repeatedly applying a single update rule until convergence:

$$V_{cj} \leftarrow V_{cj}\left(1 - \beta + \beta\frac{(\mathbf{S}\mathbf{V})_{cj}}{(\mathbf{V}\mathbf{V}^{\mathsf{T}}\mathbf{V})_{cj}}\right) \tag{3}$$

where $0 < \beta \leq 1$ is a user-defined parameter which controls the rate of convergence. We have observed that, not only is the algorithm efficient in comparison to other NMF algorithms, but it also has a tendency to produce relatively sparse factors representing localised clusters.

It has been demonstrated that supervised ensembles are most successful when constructed from a set of accurate classifiers whose errors lie in different parts of the data space [88]. Similarly, unsupervised ensemble procedures typically seek to encourage diversity with a view to improving the quality of the information available in the integration phase. A simple but effective strategy is to rely on the inherent instability of randomly-initialised factorisation algorithms. By employing a stochastic initialisation scheme, symmetric NMF will generally converge to a variety of different local solutions when applied multiple times to the same matrix $\mathbf{S}$. The level of diversity among the ensemble members can be increased by varying the number of clusters in each base clustering, such as by randomly selecting a value k_i from a predefined range $[k_{min}, k_{max}]$. An important benefit of this strategy is that it ameliorates a model selection problem with NMF which is highly sensitive to the choice of the number of basis vectors k_i.

Further improvements in performance and accuracy can be achieved by seeding each NMF factorisation using the output of the less computationally expensive kernel k-means algorithm [89]. Specifically, to seed the i-th base clustering, we randomly assign data objects to k_i clusters and apply kernel k-means to the matrix $\mathbf{S}$. The resulting disjoint clustering can be represented as an $n \times k_i$ partition matrix, where the j-th column is a binary membership indicator for the j-th cluster. This partition matrix is subsequently used as the initial factor for symmetric NMF. The use of random cluster assignment and the tendency of kernel k-means to converge to a local solution ensures that sufficient diversity in the ensemble is maintained.

A.2 Ensemble Integration Phase

We now propose an approach for combining the factors produced during the generation phase to construct a soft hierarchical clustering of the original dataset.

Graph Construction. From the generation phase, we have a collection of τ factors, giving a total of $l = (k_1 + k_2 + \cdots + k_\tau)$ individual basis vectors across all factors. We denote these vectors as the set $\mathbb{V} = \{v_1, \ldots, v_l\}$. This set can be modelled as a complete weighted graph consisting of l vertices, where each vertex represents a basis vector v_i. The weight on each edge indicates the similarity between the pair of vectors associated with the two vertices. The value of the edge weight is computed as the $[0, 1]$-normalised Pearson correlation between a pair of vectors (v_i, v_j):

$$ncor(v_i, v_j) = \frac{1}{2}\left(\frac{(v_i - \bar{v}_i)^\intercal (v_j - \bar{v}_j)}{||(v_i - \bar{v}_i)|| \cdot ||(v_j - \bar{v}_j)||} + 1\right) \tag{4}$$

The entire graph can be represented by its adjacency matrix $\mathbf{L}$, where $L_{ij} = ncor(v_i, v_j)$.

Meta-Clustering. Following the lead of the MCLA approach described by Strehl & Ghosh [21], we produce a "meta-clustering" (*i.e.* a clustering of clusters) of the graph formed from the basis vectors in $\mathbb{V}$. This is achieved by applying an agglomerative clustering algorithm to $\mathbf{L}$, resulting in a disjoint hierarchy of "meta-clusters" (*i.e.* tree nodes containing basis vectors from $\mathbb{V}$). Rather than using a traditional linkage function such as average linkage during the agglomeration process, we compute the similarity between pairs of meta-clusters based on the *min-max* graph partitioning objective [90]. This linkage function has a tendency to produce clusters which are relatively balanced in size. Formally, given the matrix $\mathbf{L}$, the min-max inter-cluster similarity between a pair of meta-clusters (M_a, M_b) is defined as:

$$sim(M_a, M_b) = \frac{s(M_a, M_b)}{s(M_a, M_a)\ s(M_b, M_b)} \tag{5}$$

such that

$$s(M_a, M_b) = \sum_{v_i \in M_a} \sum_{v_j \in M_b} L_{ij}$$

Soft Hierarchy Construction. The output of the meta-clustering procedure is a clustering of the basis vectors in $\mathbb{V}$, in the form of a traditional disjoint hierarchical tree. We wish to transform this into a *soft hierarchical clustering* of the original dataset. That is, a binary tree structure, where each node M_a in the hierarchy is associated with an n-dimensional vector y_a containing non-negative real values indicating the degree of membership for all n data objects. In practice, these node membership vectors will become increasingly sparse as we proceed further down the tree, representing more localised sub-structures.

To transform the meta-clustering into a soft hierarchy, we process each node M_a in the meta-clustering tree, computing the membership vector y_a as the mean of all the basis vectors contained in M_a:

$$y_a = \frac{1}{|M_a|} \sum_{v_i \in M_a} v_i \tag{6}$$

We associate the vector y_a with the position held by the node M_a in the original meta-clustering tree. By preserving the parent-child relations from that tree, these vectors can be linked together to form a soft hierarchy as defined above.

Final Model Selection. A hierarchical meta-clustering of the l basis vectors in $\mathbb{V}$ will yield a corresponding soft hierarchy containing l leaf nodes. However, a certain proportion of these nodes will be redundant, where the membership vectors of a pair of sibling nodes may be nearly identical to the membership vector of their parent node. This situation will arise when a tree node in the meta-clustering of $\mathbb{V}$ contains basis vectors that are highly similar to one another. Ideally we would like to prune the soft hierarchy to remove all redundant leaf and internal nodes, thereby facilitating visualisation and human interpretation.

The concept of ensemble *stability* has previously been considered as a means of identifying an appropriate cut-off point in a disjoint hierarchy [91]. Here we propose a stability-based approach to identifying an appropriate cut-off level, which is applicable to a soft hierarchy. Specifically, we consider a tree node to be *stable* if the basis vectors in the corresponding meta-cluster are highly similar, while an *unstable* node has a corresponding meta-cluster consisting of basis vector that are dissimilar to one another. To numerically assess stability, we measure the extent to which an internal node can be split into diverse sub-nodes. Given a node M_a with child nodes (M_b, M_c), this can be quantified in terms of the weighted similarity between the membership vector y_a and the pair of vectors (y_b, y_c) associated with the child nodes:

$$split(M_a) = \frac{|M_b|}{|M_a|} ncor(y_a, y_b) + \frac{|M_c|}{|M_a|} ncor(y_a, y_c) \tag{7}$$

From this, we define the *splitting factor* of an internal node M_a as the minimum value for Eqn. 7 among M_a and all child nodes below M_a in the hierarchy. A lower value indicates a lower degree of stability for the branch beginning at M_a. Using this criterion, we can prune a soft hierarchy by processing each internal node M_a in the tree, starting at the root node. The child nodes of M_a (together

Inputs:

- $\mathbf{S}$: Non-negative pairwise similarity matrix.
- τ: Number of factorisations to generate.
- $[k_{min}, k_{max}]$: Range for selecting number of clusters in each factorisation.

Generation Phase:

1. For $i = 1$ to τ
   - Randomly select $k_i \in [k_{min}, k_{max}]$.
   - Apply kernel k-means to $\mathbf{S}$ to initialise $\mathbf{V}_i \in \mathbb{R}^{n \times k_i}$.
   - Apply symmetric NMF to $\mathbf{S}$ and $\mathbf{V}_i$.
   - Add each column vector of $\mathbf{V}_i$ to the set $\mathbb{V}$.

Integration Phase:

1. Construct the adjacency matrix $\mathbf{L}$ from the set $\mathbb{V}$ according to Eqn. 4.
2. Apply min-max hierarchical clustering to $\mathbf{L}$ to produce a meta-clustering of the basis vectors.
3. Build a soft hierarchy by computing the mean vector for each tree node in the meta-clustering.
4. If required, recursively remove redundant tree nodes based on the *splitting factor* criterion.

Fig. 5. Summary of Ensemble NMF clustering algorithm

with all the nodes below them) are removed from the tree if the splitting factor of M_a is greater than or equal to a user-defined threshold λ. In practice we have observed that a threshold value of $\lambda = 0.9$ frequently leads to the elimination of redundant nodes without removing those containing informative structures.

The pruning procedure outlined above allows us to construct a tree with k leaf nodes, where the value k does not need to be specified *a priori*. As with cut-off techniques used to convert a disjoint hierarchy to a flat partition, we can produce a flat soft clustering from the leaf nodes in the tree. Specifically, we construct a $n \times k$ matrix whose columns correspond to the vectors of the k non-redundant leaf nodes in the soft hierarchy. Unlike spectral dimension reduction procedures such as PCA, standard NMF techniques do not produce an ordering of the new dimensions in terms of importance. To produce an ordering of the columns in the flat soft clustering, the related k leaf nodes may be ranked based on their *splitting factor*, with the first column corresponding to the most stable node. The complete Ensemble NMF algorithm is summarised in Figure 5.