# Comparing Non-parametric Ensemble Methods for Document Clustering

Edgar Gonzàlez and Jordi Turmo

TALP Research Center
Universitat Politècnica de Catalunya
{egonzalez,turmo}@lsi.upc.edu

**Abstract.** The biases of individual algorithms for non-parametric document clustering can lead to non-optimal solutions. Ensemble clustering methods may overcome this limitation, but have not been applied to document collections. This paper presents a comparison of strategies for non-parametric document ensemble clustering.

## 1 Introduction

As the availability of large amounts of textual information is unlimited in practice, supervised processes for mining these data can become highly expensive for human experts. For this reason, unsupervised methods are a central topic of research on tasks related to text mining. One of these tasks is document clustering. Most of the work in this area deals with parametric approaches [1, 2], in which the number of clusters has to be provided *a priori*.

On the contrary, non-parametric document clustering can be defined as the process of grouping similar documents without requiring *a priori* either the number of document categories or a careful initialization of the process from a human user. Some approaches to this task consist in repeatedly applying an iterative clustering algorithm (e.g., k-Means) to obtain a set of clusterings with a different number of clusters and starting conditions each one, and then selecting the best clustering using some model criterion [3]. Some others estimate the number of clusters *a priori* considering mathematical properties of the input documents, and then apply an iterative clustering algorithm [4]. Other approaches are based on the use of a hierarchical clustering algorithm (e.g., Hierarchical Agglomerative Clustering (HAC)) and a criterion function to select the best number of clusters in the dendrogram [5]. Recently, hybrid methods have been experimented, using the output generated from one clustering algorithm to initialize another one [6, 7].

However, each proposed approach implements some algorithm, which has an intrinsic and particular bias; uses a certain document representation; and depends on a different document similarity measure. All these assumptions lead the clustering process to a particular solution that may not be the optimal document clustering. In order to overcome this limitation, ensemble methods can

E. Kapetanios, V. Sugumaran, M. Spiliopoulou (Eds.): NLDB 2008, LNCS 5039, pp. 245–256, 2008.

be used. From a general point of view, given multiple clusterings, these methods aim at finding a combined clustering with better quality [8].

Most work in ensemble document clustering has focused on parametric approaches [9, 10, 11]. However, non-parametric ensemble approaches for generic clustering have appeared recently, such as [12].

We believe that two questions remain hence unanswered in the state of the art with respect to the use of ensemble methods for document clustering:

- **How well do ensemble methods perform for non-parametric document clustering?** Non-parametric methods have not been tested thoroughly on document collections so far.
- **How well do different individual clustering strategies perform in the context of non-parametric ensemble document clustering?** The influence of the strategy used to find individual clusterings to be later combined has often been overlooked. Different strategies need to be compared.

This paper deals with both questions. It evaluates non-parametric clustering algorithms on document collections; and it presents an empirical comparison of the effectiveness of two different strategies for the generation of clustering ensembles: one relying on massive randomization of a single algorithm, and another relying on few but heterogeneous different algorithms.

The rest of the paper is organized as follows: Section 2 settles the problem of non-parametric document ensemble clustering. Sections 3 and 4 describe the two considered generation strategies for the clustering ensembles. Section 5 then gives an overview of the experiments performed and their results. Last, Section 6 draws conclusions of our work.

## 2 Non-parametric Document Ensemble Clustering

Having $\mathcal{D} = \{d_1 \dots d_n\}$ a set of documents, a clustering, $\Pi$, of this set is a partition of $\mathcal{D}$ into a set, $\{\pi_1 \dots \pi_k\}$, of $k$ disjoint clusters, $\pi_i$. The clustering, $\Pi$, can also be viewed as a function mapping documents, $d_l$, onto labels $\{1 \dots k\}$ corresponding to clusters $\{\pi_1 \dots \pi_k\}$, where $\Pi(d_l) = i \leftrightarrow d_l \in \pi_i$.

Bearing this in mind, the aim of clustering combination is to find a clustering, $\bar{\Pi}$, which is the consensus of $r$ clusterings, $\{\Pi_1 \dots \Pi_r\}$, by means of a consensus function $\Gamma$.

Two settings are classically considered for this problem, according to whether the consensus function accesses or not the original representation of the data. It is usual to refer to the case when the original data are not accessed as *cluster ensemble* [9]. This setting allows combination of clusterings obtained using different document representations. We stick to it in this paper, as it is a more general framework than the former and, in addition, it is widely used by the machine learning research community [8, 13, 14].

For our experiments, we have focused on the non-parametric ensemble clustering approach of [12], which includes methods for the determination of the number of clusters. Among the methods proposed in the paper, we have chosen

**Algorithm 1.** Major ensemble strategy

**Parameter:** $\mathcal{D}$ a document collection
**Parameter:** $r$ a natural number
**Parameter:** $k_{max}$ a natural number
**Parameter:** $\varphi$ a supervised clustering algorithm

1: **for** $j = 1 \dots r$ **do**
2: Select a number of clusters at random $k_j \in \{2 \dots k_{max}\}$
3: Select $k_j$ documents at random as starting centroids
4: Apply $\varphi$ to $\mathcal{D}$ to obtain clustering $\Pi_j$
5: **end for**
6: Return ensemble $\{\Pi_1 \dots \Pi_r\}$

the AGGLOMERATIVE algorithm, enhanced with LOCALSEARCH. This combination was found in preliminary experiments to outperform the rest of the proposed approaches on the evaluation data collections[1].

## 3 Major Ensemble Strategy

There has been recent interest in research on ensemble clusterings from repeated runs of randomly initialized algorithms [8, 13]. In these works, the results obtained were competitive to other proposed approaches for a variety of classical clustering problems in machine learning.

For this reason, as a first strategy we have considered repeatedly applying a single individual clustering method a high number of times, with different starting conditions selected at random. The main properties of this strategy are the following:

- The resulting clusterings share the same data representation.
- The algorithm is unique, hence, the implicit bias introduced by the clustering process is always the same.
- The size of the ensemble can be high.

The procedure is detailed in Algorithm 1. First, a number of clusters $k$ from 2 to $k_{max}$ is selected at random. Then, $k$ documents are selected at random from the collection, and are given as starting centroids to a clustering algorithm, $\varphi$. This process is repeated a number of times $r$, and the $r$ resulting clusterings are combined using the ensemble clustering function.

The parametric clustering algorithm, $\varphi$, is a parameter of the method. For our experiments, we have used the EM-based clustering algorithm of [15]. This algorithm has obtained competitive results for text classification, and has already been used for document clustering [7]. Other parameters that need to be chosen are the number of individual clusterings, $r$, and the maximum number of clusters,

[1] Further details about these algorithms can be found on the original paper.

$k_{max}$. For the considered document collections, the best results among the set of explored parameter values were obtained with $r = 50$ and $k_{max} = 10$.

We will refer to this method as **Major**.

## 4 Minor Ensemble Strategy

Whereas the **Major** combination strategy we have described in the previous section is based on the repetition of a randomly initialized single clustering algorithm, the second strategy we have considered, **Minor**, is based on the use of a small number of heterogeneous, unsupervised and deterministic clustering algorithms. As in this case there is no randomization, it is crucial to the method that the biases provided by the algorithms be substantially different from each other. For this reason we have selected the following unsupervised clustering methods:

- A classical method, consisting of a hierarchical algorithm and a clustering criterion function.
- A hierarchical-iterative hybrid method. Specifically, the hybrid method of [7], which has been shown to give good performance for unsupervised document clustering of different real-world collections.
- A new version of the previous hybrid method, based on information theory, which we have devised specially for this purpose.

A description of each one of them follows.

### 4.1 Hierarchical Method

In order to generate a dendrogram, the Agglomerative Information Bottleneck algorithm (aIB) is used. [16] applies the algorithm to a variety of standard supervised clustering problems. Particularly, aIB showed good performance for the task of supervised document clustering.

After the dendrogram is built, the Calinski and Harabasz C score [17] is used to determine the level of the dendrogram at which the best clustering occurs. This score has been compared to other similar criteria to determine the number of clusters in a collection, and shown to be the most efficient one [3]. Its value is the normalized ratio of the inter-cluster distances (between documents of different clusters) against intra-cluster distances (within documents of the same cluster). The level at which this value is highest is selected as the best estimation of the number of clusters.

We will refer to this method as **Hi**.

### 4.2 Geometric Hybrid Method

The method presented in [7] tries to find a good initial clustering for an iterative refinement algorithm. Iterative refinement algorithms are known to be efficient

and give good performance, but nevertheless are sensitive to the choice of the initial model, and require the number of clusters to be provided. In particular, a good estimation of the number of clusters is mandatory for a good initial model, even if this model does not cover all documents in the collection.

An outline of the procedure follows:

1. A hierarchical algorithm is used to find a dendrogram.
2. The inner nodes in the dendrogram are scored according to different heuristics, based in minimizing the distances within documents covered by the node, and maximizing distances to the rest of the documents[2].
3. The nodes the best scored according to the heuristics are chosen as clusters for an initial clustering candidate. A different candidate is built for each heuristic.
4. These candidates are scored using a global quality function, and the best scored candidate is selected.
5. This candidate is used as initial model for an iterative refinement algorithm, to produce a final clustering solution.

In its original implementation, the method is specified using a geometric point of view:

- Documents are represented as $tf \cdot idf$ vectors of words.
- The distance metric is cosine distance.
- The hierarchical algorithm used is HAC with group average distance as distance between clusters, which was pointed as the most suitable distance in HAC context by published evaluations of the algorithm [1].
- The global quality function is Calinski and Harabasz C score.
- The iterative refinement algorithm applied is the EM-based algorithm of [15].

We will refer to this method as **Geo**.

### 4.3 Information Theoretical Hybrid Method

Even if geometric clustering methods remain the state of the art, there has been a recent interest in applying information theoretical measures to the task of document clustering [16, 18]. Following this general direction of research, and to find a view of the data different from that of **Geo**, we have made a new version of the aforementioned hybrid method using information theoretical concepts:

- Documents are represented as conditional probability distributions of words.
- The distance metric is Jensen-Shannon divergence. There are other measures coming from information theory that could be useful to define a document distance, such as Kullback-Leibler divergence or mutual information. However, on the contrary of Jensen-Shannon divergence, they are not symmetric or require absolute continuity.

[2] For simplicity, the details about these heuristics have been elided in this paper.

- The hierarchical algorithm used is aIB.
- The global quality function used is a specially devised Message Length Criterion, described below in Section 4.3.
- The iterative refinement algorithm applied is Divisive Information Theoretical Clustering (DITC) [18]. This algorithm includes devices to deal with sparseness and high dimensionality of data, and was shown to give good performance on document collections.

We will refer to this method as **IT**.

**Message Length Criterion.** Classical information theoretical selection criteria, such as Minimum Description Length or Minimum Message Length, require a probability distribution, which cannot be directly derived from the dendrogram. However, we have devised a criterion to select the best clustering in the same spirit, based in coding, messages and lengths.

The idea is to use the information in a clustering $\Pi$ to send a collection of documents $\mathcal{D}$ as a message. We first send the send the centroid of each cluster using a code based on the meta-centroid of the collection (a first message of length $L_C(\Pi)$), and then send the distribution of words in each document using a code based of the centroid of the cluster to which it belongs (a second message of length $L_D(\Pi)$). Using formulae from Information Theory, the total length of this message, $L(\Pi)$, is roughly:

$$
\begin{aligned}
L(\Pi) &\approx L_C(\Pi) + L_D(\Pi) \\
L_C(\Pi) &\approx - \sum_{\substack{\pi_i \in \Pi \\ w}} p(w|c_i) \cdot \log p(w|mc) \\
L_D(\Pi) &\approx - \sum_{\substack{\pi_i \in \Pi \\ d_l \in \pi_i \\ w}} p(w|d_l) \cdot \log p(w|c_i)
\end{aligned}
$$

where $w$ are words, $c_i$ are the cluster centroids and $mc$ is the meta-centroid.

We expect *better* clusterings (i.e. more suited to the data) to allow better compression of the data and hence, shorter messages. Therefore, we select the clustering $\Pi$ which has the lowest $L(\Pi)$, expecting it to be the *best*.

This formula was the one to give the best results in preliminary experiments, compared to a version of the C score using Jensen-Shannon divergence.

Moreover, this formula was appealing to us because it includes an implicit measure of the goodness of the number of clusters (more clusters imply largest $L_C(\Pi)$ but smallest $L_D(\Pi)$, and vice versa).

## 5 Experiments

In order to evaluate and compare the performance of the two proposed ensemble strategies, **Major** and **Minor**, between them and to individual clustering

approaches (of which **Geo** can be considered a baseline in the state of the art), we have carried out a series of experiments. The following sections explain the experimental framework, and present their results.

### 5.1 Evaluation Data

Six different real-world English document collections have been used in our experiments:

**APW.** The Associated Press (year 1999) subset of the AQUAINT collection. Due to memory limitations in our test machines, the collection was reduced to the first 5000 documents.

**EFE.** A collection of news-wire documents from year 2000 provided by the EFE news agency.

**LAT** The Los Angeles Times subset of the TREC-5 collection. For the same reason as in APW, again only the first 5000 documents were selected.

**REU.** A subset of the Reuters-21578 text categorization collection, which includes only the ten most frequent categories. Similarly to previous work, we use the ModApte split [7, 15], but, since our algorithms are unsupervised, we use the test partition directly.

**SMT.** A collection previously developed and used for the evaluation of the SMART information retrieval system.

**SWB.** A subset of the Switchboard conversational speech corpus, which contains the 22 topics which were treated in more than fifty conversations. Each side of the conversation was considered a separate document.

Following other research work [2, 7], the documents were pre-processed by discarding stop words and numbers, converting all words to lower case, and removing terms occurring in a single document. Table 1 lists relevant collection characteristics after pre-processing (number of documents, categories and terms).

**Table 1.** Evaluation data sets

| Collection | Docs | Cats | Terms |
|---|---|---|---|
| APW | 5000 | 11 | 27366 |
| EFE | 1979 | 6 | 10334 |
| LAT | 5000 | 8 | 31960 |
| REU | 2545 | 10 | 6734 |
| SMT | 5467 | 4 | 11950 |
| SWB | 2682 | 22 | 11565 |

### 5.2 Evaluation Metrics

The quality of the clustering solutions is measured using the metrics of purity, inverse purity and $F_1$. These metrics have been widely used to evaluate the

performance of document clustering algorithms [2], and are based in comparing the clustering to a partition which is considered *true*.

If we have a partition of the documents in $\mathcal{D}$ into a set of disjoint categories considered *true*, these metrics can be defined as:

**Pur.** Purity evaluates the degree to which each cluster contains documents from a single category. The purity of a cluster is the fraction of the documents in the cluster that belong to its majoritarian category. The overall purity is the average of all cluster purities, weighted by cluster size.

**IPur.** Inverse purity evaluates the degree to which the documents in a category are grouped in a single cluster. The inverse purity of a category is the fraction of the documents in the category that are assigned to its majoritarian cluster. The overall inverse purity is the average of all category inverse purities, weighted by category size.

**$\mathbf{F_1}$.** $F_1$ is a global performance score, and is calculated as the harmonic mean of purity and inverse purity.

### 5.3 Experimental Setup

Each collection was clustered using each of the proposed methods. For the **Geo**, **Hi**, **IT** and **Minor** methods, a single run was performed, as these methods are deterministic.

For the **Major** method, we performed five runs and the results presented are the average of all the runs. As mentioned in Section 3, the results are those obtained with $r = 50$ and $k_{max} = 10$, which were the parameter values to provide the best $F_1$ scores in average across all collections.

### 5.4 Results

Tables 2, 3 and 4 show the results obtained by each method in each collection. For each collection, the best results are highlighted.

In addition, Table 5 shows the number of clusters $k$ estimated by each method in each collection. We include two numbers for each method, the total number of clusters (**All**), and the number of *relevant* clusters (**Rel**). The reason for this is that we have found that the AGGLOMERATIVE algorithm tends to find a high number of clusters, but many of them are small, possibly corresponding to outliers among the data.

Given that these small clusters are not relevant to the evaluation (and their detection as outliers is, in fact, an advantageous byproduct of the method), to obtain a more useful measure we have filtered those clusters smaller than a fourth of the average category size in the collection. The remaining ones are considered *relevant*, and their number is the figure appearing in the table. The number of categories (**Cats**) in each collection is also included in the table.

Following sections discuss the obtained results.

**Overall Comparison.** It can be seen how the **Major** approach outperforms the rest of the approaches in almost all collections in terms of $F_1$, and is also the

**Table 2.** $F_1$ values for all methods and collections

| | Geo | Hi | IT | Major | Minor |
|---|---|---|---|---|---|
| APW | 0.75 | 0.74 | 0.63 | 0.75 | 0.72 |
| EFE | 0.61 | 0.61 | 0.58 | 0.62 | 0.60 |
| LAT | 0.67 | 0.67 | 0.67 | 0.75 | 0.67 |
| REU | 0.88 | 0.79 | 0.76 | 0.88 | 0.88 |
| SMT | 0.85 | 0.82 | 0.71 | 0.93 | 0.91 |
| SWB | 0.79 | 0.26 | 0.53 | 0.44 | 0.66 |

**Table 3.** Purity values for all methods and collections

| | Geo | Hi | IT | Major | Minor |
|---|---|---|---|---|---|
| APW | 0.78 | 0.63 | 0.72 | 0.80 | 0.74 |
| EFE | 0.73 | 0.60 | 0.64 | 0.75 | 0.70 |
| LAT | 0.78 | 0.66 | 0.75 | 0.73 | 0.79 |
| REU | 0.84 | 0.73 | 0.77 | 0.86 | 0.85 |
| SMT | 0.92 | 0.71 | 0.89 | 0.93 | 0.93 |
| SWB | 0.69 | 0.15 | 0.38 | 0.29 | 0.53 |

**Table 4.** Inverse purity values for all methods and collections

| | Geo | Hi | IT | Major | Minor |
|---|---|---|---|---|---|
| APW | 0.73 | 0.88 | 0.56 | 0.70 | 0.70 |
| EFE | 0.52 | 0.63 | 0.53 | 0.53 | 0.53 |
| LAT | 0.59 | 0.68 | 0.61 | 0.79 | 0.59 |
| REU | 0.92 | 0.86 | 0.76 | 0.90 | 0.89 |
| SMT | 0.80 | 0.97 | 0.58 | 0.92 | 0.90 |
| SWB | 0.94 | 0.92 | 0.91 | 0.97 | 0.89 |

**Table 5.** Number of clusters $k$ for all methods and collections

| | | Geo | | Hi | | IT | | Major | | Minor | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cats | All | Rel | All | Rel | All | Rel | All | Rel | All | Rel |
| APW | 11 | 10 | 9 | 3 | 3 | 8 | 8 | 60.6 | 7.0 | 19 | 7 |
| EFE | 6 | 12 | 7 | 4 | 4 | 5 | 5 | 69.0 | 6.2 | 14 | 7 |
| LAT | 8 | 14 | 9 | 6 | 6 | 7 | 7 | 27.2 | 4.8 | 40 | 7 |
| REU | 10 | 6 | 6 | 4 | 4 | 6 | 6 | 18.2 | 5.2 | 13 | 6 |
| SMT | 4 | 6 | 5 | 3 | 3 | 9 | 7 | 20.6 | 4.0 | 18 | 4 |
| SWB | 22 | 15 | 15 | 3 | 3 | 8 | 8 | 10.4 | 5.8 | 22 | 12 |

best approach in terms of purity in four of the six collections. Its performance in terms of inverse purity is not always the best, but it is always comparable to that of the rest of the methods.

The performance of **Minor** and **Geo** is quite similar in terms of purity, but **Minor** suffers from lower inverse purity, so overall its $F_1$ is also lower. The **Hi**

method usually gives solutions with a high inverse purity but a low purity, so in many cases the global $F_1$ scores are lower than other approaches. Lastly, the results of **IT** do not stand out in any aspect, and its utility outside the **Minor** combination seems limited, at least at the light of these results.

Nevertheless, we have applied a Friedman test, followed by pairwise Nemenyi tests, to account for statistical significance of these differences [19]. We only found that **Hi** is worse than **Major**, **Minor** and **Geo** in terms of purity; and that **IT** is worse than **Major** in terms of $F_1$. No other significant differences were found. This is relevant, because it means there is no empirical evidence supporting the rejection of any of the **Geo**, **Major** or **Minor** methods as less suitable to the task than the others, in terms of purity, inverse purity or $F_1$ score.

**Estimation of the Number of Clusters.** Concerning the estimated number of clusters, we can see how the ensemble-based approaches greatly overestimate the total number of clusters (**All**). As explained in Section 5.4, this is caused by the presence of a large number of small clusters, and the figures for the number of relevant clusters (**Rel**) are much closer to the actual number of categories (**Cats**).

However, it can be seen that the estimation of the total number of clusters by **Minor** is more accurate than that by **Major** in all but the LAT collection. **Major** shows a bias for purity, and shows a slightly displeasing tendency to disgregation.

Regarding the individual methods, whereas the estimation by **Geo** and **IT** is fairly accurate; **Hi** shows a tendency to underestimation, which explains its high inverse purity values and low purity values. The individual methods do not present such a large number of small clusters, which on the one hand means there is not such a risk of disgregation, but on the other one can mean a more limited capability to detect outliers.

**Minor Method.** As mentioned before, the performance of **Minor** method is only significantly better than that of **Hi** in terms of purity. Nevertheless, the results of the combination seem comparable to those of **Geo**, and better than those of **IT**.

Overall, **Minor** offers a greater stability across document collections than its components **Hi** and **IT**. Moreover, the fact that neither **Hi** nor **IT** do not perform competitively on document collections (particularly on SWB) suggests that using some other algorithm more suitable for this kind of data the performance of **Minor** could be boosted, and more competitive results could be obtained.

For this reason, together with the facts that its performance is not significantly worse than that of **Major**; that it gives a better estimation of the number of clusters; and that it has no parameters needing to be tuned, whereas **Major** requires the values of $k_{max}$ and $r$ have to be determined (see Section 3); we believe that the **Minor** method remains an attractive approach, and that more research should be carried on the topic of small ensembles of heterogeneous clusterings.

**SWB Collection.** The main exception to the general behaviour seems to be the SWB collection. Almost all methods experiment a considerable decrease in purity when applied to this data set. We believe this comes from the fact that, the size of all categories in SWB is quite similar, whereas for the rest of collections a few large categories cover most of the documents. This makes the SWB collection harder than the rest, and specially sensitive to underestimation of the number of clusters.

The fact that all the considered methods do underestimate the number of clusters (as can be seen in the **Rel** columns of Table 5), causes low values of purity (in some causes dramatically low, e.g. **Hi**), and hence of $F_1$. Only **Geo** and, to a lesser extent, **Minor** seem able to find a reasonable (even if still underestimated) number of relevant clusters (column **Rel**) in this collection.

## 6 Conclusions

We have studied the application of a non-parametric ensemble clustering approach to document collections, and considered two different strategies for the generation of the clustering ensembles. Lastly, we have carried a set of experiments with real-world data.

At the light of the results, we can conclude that non-parametric ensemble methods do perform competitively for clustering of document collections. Regarding the two considered strategies, whereas the **Major** approach gives better figures of purity and $F_1$ score, the differences with **Minor** are not statistically significant, its estimation of the number of clusters is worse, and it has a number of parameters to be tuned.

For these reasons, and because there is further room for improvement of the individual components of **Minor**, we believe that the results of this heterogeneous approach can be boosted, and that it remains an attractive approach for the task.

## Acknowledgments

This work has been partially funded by the Spanish Text-Mess Project (TIN2006-15265-C06); the Commissionate for Universities and Research of the Department of Innovation, Universities and Enterprises of the Catalan Government; and the European Social Fund.

## References

1. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In: Proc. of CIKM (2002)
2. Zhao, Y., Karypis, G.: Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning 55(3) (2004)
3. Milligan, G.W., Cooper, M.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrica 50 (1985)

4. Li, T., Ma, S., Ogihara, M.: Document clustering via adaptive subspace iteration. In: Proc. of SIGIR (2004)
5. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society Series B 63(2) (2001)
6. Fraley, C., Raftery, A.: How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41(8) (1998)
7. Surdeanu, M., Turmo, J., Ageno, A.: A hybrid unsupervised approach for document clustering. In: Proc. of KDD (2005)
8. Topchy, A., Jain, A.K., Punch, W.: Clustering ensembles: Models of consensus and weak partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12) (2005)
9. Strehl, A., Ghosh, J.: Cluster ensembles - A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3 (2002)
10. Siersdorfer, S., Sizov, S.: Restrictive clustering and metaclustering for self-organizing document collections. In: Proc. of SIGIR (2004)
11. Greene, D., Cunningham, P.: Efficient ensemble methods for document clustering. Technical report, Department of Computer Science, Trinity College Dublin (2006)
12. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. In: Proc. of ICDE (2005)
13. Fred, A., Jain, A.: Robust data clustering. In: Proc. of CVPR (2003)
14. Li, T., Ogihara, M., Ma, S.: On combining multiple clusterings. In: Proc. of CIKM (2004)
15. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39(2/3) (2000)
16. Slonim, N.: The Information Bottleneck: Theory and Applications. PhD thesis, The Hebrew University (2003)
17. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics 3 (1974)
18. Dhillon, I., Guan, Y.: Information theoretic clustering of sparse co-occurrence data. In: Proc. of ICDM (2003)
19. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7 (2006)