Interactive Clustering of Text Collections According to a User-Specified Criterion

Ron Bekkerman
University of Massachusetts
Amherst MA, USA
ronb@cs.umass.edu

Hema Raghavan
University of Massachusetts
Amherst MA, USA
hema@cs.umass.edu

James Allan
University of Massachusetts
Amherst MA, USA
allan@cs.umass.edu

Koji Eguchi
National Institute of Informatics
Tokyo, Japan
eguchi@nii.ac.jp

Abstract

Document clustering is traditionally tackled from the perspective of grouping documents that are *topically* similar. However, many other criteria for clustering documents can be considered: for example, documents' genre or the author's mood. We propose an interactive scheme for clustering document collections, based on any criterion of the user's preference. The user holds an active position in the clustering process: first, she chooses the types of features suitable to the underlying task, leading to a task-specific document representation. She can then provide examples of features—if such examples are emerging, e.g., when clustering by the author's sentiment, words like 'perfect', 'mediocre', 'awful' are intuitively good features. The algorithm proceeds iteratively, and the user can fix errors made by the clustering system at the end of each iteration. Such an *interactive clustering* method demonstrates excellent results on clustering by sentiment, substantially outperforming an SVM trained on a large amount of labeled data. Even if features are not provided because they are not intuitively obvious to the user—e.g., what would be good features for clustering by genre using part-of-speech trigrams?—our multi-modal clustering method performs significantly better than k-means and Latent Dirichlet Allocation (LDA).

1 Introduction

The problem of data clustering is generally underspecified unless criteria for clustering are explicitly provided. For example, given a set of objects of various colors and shapes, it is unclear whether clustering should be performed according to the objects color, shape, or both. In the absence of labeled instances, a clustering criterion might be expressed in terms of the data representation: e.g., if only shapes of objects are known, there is no more doubt about the clustering criterion.

When we talk about clustering text documents, we usually assume that the clustering will be by topic and we typically approach it using a Bag-Of-Words (BOW) representation that ignores word order [Willett, 1988]. However, there is no reason that text documents must be clustered in that way: there are numerous non-topical criteria that could be considered—e.g., clustering by sentiment [Turney, 2002], style, genre, author's mood, and so on. Other criteria may be esoteric or application-specific: e.g., clustering by the author's age, by the age of the documents, by their credibility, expressiveness, readability, etc. It is unlikely that a simple BOW representation will be sufficient for all of those purposes, meaning that most will require specific document representations. Intuitively, some of these representations will be based primarily on syntax, while others are likely to have a semantic nature.

This study proposes a unified framework for clustering document collections according to nearly any criterion of the users choice. (We restrict ourselves to hard clustering—i.e., partitioning—of a document collection.) The user is first asked to choose *types of features* suitable for clustering by the desired criterion. For example, genres may be represented by sequences of Part-Of-Speech (POS) tags, by a particular focus on punctuation, stopwords, as well as by general words as captured in the standard BOW representation. The user is next asked to provide a few examples of features (*seed features*) of the chosen types, if such examples are intuitive and can be obtained without much effort—e.g., when clustering by authors mood, words like 'angry', 'happy', 'upset' might be easily suggested.

The clustering system then represents documents based on the users choice and applies a multi-modal clustering method [Bekkerman and Sahami, 2006]. When seed features are provided, the system iteratively clusters documents represented over the chosen features and then enriches feature sets with other useful features. The user can choose to intervene (or not) after each iteration, in order to fix possible mistakes made by the system on the feature level (no document labeling is required).

We illustrate the effectiveness of our approach on two domains: clustering by genre and clustering by author's sentiment. Genre is a type of a domain where providing examples of features is non-trivial: it is not intuitive, e.g., whether noun phrases are more effective than verbs. Sentiment classification is one where words like 'brilliant' and so on are easily recognizable as useful, when using BOW features.

There has been work on interactive topical clustering where the user corrects clustering errors on a *document* basis [Basu *et al.*, 2004], but that effort is more time consuming than feedback on features [Raghavan *et al.*, 2005]. Other re-

cent work has had the user select important keywords for (supervised) categorization, thereby leveraging the user's prior knowledge [Dayanik *et al.*, 2006; Raghavan *et al.*, 2005]—approaches that are more like that of our framework. Raghavan *et al.* [2005] further support this direction in the finding that users can identify useful features with reasonable accuracy as compared to an oracle. Liu *et al.* [2004] experiment with labeling words instead of documents for text classification, providing the user with a list of candidate words from which to select potentially good seed words, based on which a training set is constructed from a set of unlabeled documents. A classifier is then constructed given this training set. Liu *et al.*'s document representation is the standard BOW, which has strong topical flavor, and therefore cannot be used for clustering by any criterion (for example, our preliminary experiments show that BOW is not appropriate for clustering by author's mood). In addition, Liu *et al.*'s method involves the user only at the initial step (selecting seed words), limiting the user's control of the classification process.

In summary, we propose a new interactive learning framework for clustering by user-determined criteria (Section 2). Our multi-modal clustering method based on combinatorial MRFs (Section 3) neatly incorporates multiple feature types as well as user prior knowledge into clustering presented as a combinatorial optimization problem. We demonstrate the effectiveness of our system by testing it on genre clustering (Section 5) and multi-class clustering by author's sentiment when seed features are provided (Section 6). To our knowledge, this study is the first in which clustering (as opposed to classification) by genre is discussed and the first to perform multi-class clustering of documents by sentiment. We show that our interactive clustering outperforms state-of-the-art methods (SVM and LDA) on real-world data collections.

2 Interactive clustering scenario

We provide a step-by-step recipe for clustering documents by a particular criterion that the user has in mind:

1. Specify the number of clusters: Learning the natural number of clusters still remains an open problem. We do not attempt to solve it in this paper, instead the user is asked to specify the desired number of clusters.

2. Specify feature types: A list of various *feature types* is provided to the user. Examples of such types are: bag of words or word n-grams, POS tags or POS tag n-grams, punctuation, parse subtrees and other types of syntactic and semantic patterns that can be extracted from text. Such a list can hypothetically include a large variety of feature types that would respond to everyone's needs. From this list the user is asked to choose one or more types that best serve the particular clustering criterion.

3. Give examples of features: For each feature type chosen, the user should attempt to construct (small) sets of seed features that correspond to each category of documents. Sometimes this task is easy: e.g., if the clustering criterion is authors' sentiments, then words such as 'excellent', 'brilliant' etc. would correspond to the category of positive documents, while 'terrible', 'awful' etc. would correspond to the negative category. However, when such sets cannot be easily constructed (e.g. it is non-trivial to come up with good feature examples for clustering by genre—see Section 5), the user can skip this step and go to 4.

4. Default clustering: If m feature types are chosen, but no seed features are provided by the user, documents are represented as m distributions, each of which is over the (entire) feature sets of the corresponding type and then *multi-modal distributional clustering* [Bekkerman and Sahami, 2006] is applied.

5. Interactive Clustering: For the cases when the user has provided seed features for some of the feature types, we propose a new model for multi-modal clustering, which combines regular clustering of non-seeded variables with an incremental, bootstrapping procedure for seeded variables:

1. Represent documents as distributions over the sets of seed features. Ignore documents with zero probability given the seed features. Cluster the remaining documents using the distributional clustering method.
2. Stop if most documents have been clustered (see Section 6 for details).
3. Represent *all* features of the *clustered* documents as distributions over the document clusters. Ignore features that have zero probability given the clustered documents. Cluster the remaining features using the distributional clustering method.
4. Select feature clusters that contain the original seed words. Let the user revise the selected clusters: noisy features can be deleted; misplaced features can be relocated; new features can be added. The revised clusters of features are the new sets of seed features. Go to 5.1.

3 Combinatorial MRFs for clustering

A *combinatorial Markov random field (Comraf)* [Bekkerman and Sahami, 2006] is a new framework for multi-modal learning in general, and for multi-modal clustering in particular. Multi-modal (hard) clustering is a problem of simultaneously constructing m partitionings of m data modalities, e.g. of documents, their words, authors, titles etc. While clustering modalities simultaneously, one would overcome the statistical sparseness of the data representation, leading to a dense, smoothed joint distribution of the modalities that would result in (hypothetically) more accurate clusterings than the ones obtained when each modality is clustered separately. Bekkerman *et al.* [2005] empirically justify this hypothesis.

A Comraf model for multi-modal clustering is an undirected graphical model in which each data modality $\mathcal{X}_i : 1 \leq i \leq m$ corresponds to *one* discrete random variable (r.v.). This r.v. is defined over *all possible* clusterings of $\mathcal{X}_i$, which implies that the support of this r.v. is exponentially large in the size of $\mathcal{X}_i$. We call such an r.v. a *combinatorial* r.v. Let X_i be an r.v. with an empirical distribution over $\mathcal{X}_i$ (e.g. over documents in the dataset); let $\tilde{X}_{ij}$ be an r.v. defined over clusters in the j-th clustering of $\mathcal{X}_i$; let $\tilde{X}_i^c$ be a combinatorial r.v. defined over all the possible clusterings of $\mathcal{X}_i$. Edges $e_{ii'}$ in the Comraf graph G correspond to interactions between modalities (the graph is not necessarily fully connected). Examples of Comraf graphs are shown in Figure 1.

The objective is to construct clusterings of modalities (or, in other words, to find values of combinatorial r.v.'s) such that the sum of pairwise Mutual Information between the clusterings of the interacting modalities is maximized:

$$\tilde{\mathbf{x}}^{\mathbf{c}*} = \arg\max_{\tilde{\mathbf{x}}^{\mathbf{c}}_j} \sum_{e_{ii'} \in \mathbf{E}} I(\tilde{X}_{ij}; \tilde{X}_{i'j}). \quad (1)$$

This objective function naturally factorizes over G, so that an efficient inference algorithm (such as Iterative Conditional Mode—ICM [Besag, 1986]) can be directly applied. The ICM algorithm iterates over each node in G, which is optimized with respect to the current values of its neighbors.

In the Comraf case, the optimization of each node is a resource-consuming process. Each clustering $\tilde{x}^c_{ij}$ can be represented as a point $(c_{j1}, c_{j2}, \ldots, c_{jn_i})$ in an n_i-dimensional hypercube H_i of all the possible clusterings (where n_i is the number of elements of the i-th modality), meaning that element 1 belongs to cluster c_{j1}, element 2 belongs to cluster c_{j2} etc. We apply the simplest combinatorial optimization algorithm—*hill climbing*, where the procedure starts at some point on H_i and greedily searches for a nearby point that satisfies Equation (1). Since the problem is non-convex, random restarts are used to overcome local optima.

In this paper, we propose an interactive learning approach, in which the user assists the clustering algorithm to avoid local optima. First, by selecting seed features, the user specifies a potentially good starting point on the hypercube H_i. Second, by correcting the constructed clustering after each iteration, the user causes a controlled jump from one region of H_i to another, in which potentially better clusterings are located.

4 Evaluation methodology

In this paper we use *clustering accuracy* as a quality measure of document clustering. Let T be the set of ground truth categories. For each cluster $\tilde{d}$, let $\gamma_T(\tilde{d})$ be the maximal number of elements of $\tilde{d}$ that belong to one category. Then, the precision $Prec(\tilde{d}, T)$ of $\tilde{d}$ with respect to T, is defined as $Prec(\tilde{d}, T) = \gamma_T(\tilde{d})/|\tilde{d}|$. The micro-averaged precision of the entire clustering $\tilde{d}^c$ is: $Prec(\tilde{d}^c, T) = \frac{\sum_{\tilde{d}} \gamma_T(\tilde{d})}{\sum_{\tilde{d}} |\tilde{d}|}$, which is the portion of documents appearing in the dominant categories. For all our experiments we fix the number of clusters to be equal to the number of categories. In this case, $Prec(\tilde{d}, T)$ equals clustering accuracy.

In our experiment with clustering by sentiment, we compare Comraf clustering results with SVM classification results. Bekkerman and Sahami [2006] show that the clustering accuracy can be directly compared with the (standard) classification accuracy if a constructed clustering is *well-balanced*, meaning that each category prevails exactly in one cluster. It appears that all our clusterings obtained using the Comraf model are well-balanced.

5 Clustering by genre

According to the scenario proposed in Section 2, let us set up an experiment of clustering documents by their genre. After fixing the number of clusters to be equal to the number of categories in our dataset (see Section 4), we decide about feature types which would best match the task of clustering by genre. Documents are labeled with genres on the basis of external criteria such as intended audience, purpose and activity type [Lee, 2001]. The notion of genre can be described in terms of the syntax/semantics duality of text: documents of different genres use different syntactic constructions and/or different vocabulary. It is not obvious whether syntactic or semantic features play a major role in clustering documents by genre. We propose to take advantage of both. We represent documents over two sets of features: words (that correspond to documents' vocabularies) and Part-Of-Speech (POS) n-grams (that correspond to the syntactic structure of text). POS n-grams are extracted from sentences in an incremental manner: the first n-gram starts with the POS tag of the first word in the sentence, the second one starts with the tag of the second word etc.

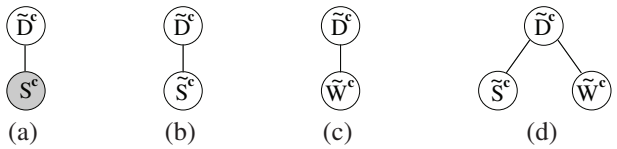

Figure 1: Comraf graphs for: (a) 1-way document clustering with POS unigrams as an *observed* r.v. (shaded node); (b) 2-way clustering of documents and POS bigrams (same as for POS 3-grams or 4-grams); (c) 2-way clustering with BOW; (d) 3-way clustering with POS bigrams and BOW.

Intuitively, one cannot come up with *particular* features that best capture documents' genres (e.g. it is hard to say whether a word 'clouds' is more often used in fiction, poetry or weather reports). To the contrary, document distributions over the *entire* set of features would be different for documents of different genres and are then the most appropriate representation of documents for clustering by genre. Thus, we apply the multi-modal clustering method described in Section 3, without the interactive learning component.

Given a document collection, let D be a random variable over its documents, W be a random variable over its words, and S be a random variable over the POS n-grams of its words. We apply a multi-modal Comraf model (Section 3) for constructing a clustering $\tilde{D}$ of documents, a clustering $\tilde{W}$ of words and/or a clustering $\tilde{S}$ of POS n-grams, by maximizing the objective from Equation (1). In this paper, we consider four Comraf models for clustering by genre:

1. POS unigrams: Since the number of POS tags in any tagging system is relatively small, it makes no sense to cluster POS unigrams. Therefore, we apply a 1-way model for clustering documents using the Comraf graph shown in Figure 1(a). The objective function from Equation (1) in this simple case has the form of $I(\tilde{D}; S)$.

2. POS n-grams, where $\mathbf{n > 1}$. The number of unique POS n-grams of order higher than 1 is exponential in n, so clustering them would be necessary. We perform a 2-way clustering with the Comraf graph from Figure 1(b) and the objective $I(\tilde{D}; \tilde{S})$.

3. Bag-Of-Words: The number of unique words in our

Doc representation	k-means	LDA	Comraf
Bag-Of-Words	9.1%	$55.4 \pm 0.1\%$	$55.7 \pm 0.2\%$
POS bigrams	23.2%	$44.7 \pm 0.2\%$	$51.0 \pm 0.2\%$
BOW + POS bigr	n/a	n/a	$\mathbf{58.5 \pm 0.6}\%$

Table 1: **Clustering by genre.** Clustering accuracy on the BNC corpus, averaged over four independent runs. Standard error of the mean is shown after the $\pm$ sign. Comraf results with other POS tuples, besides bigrams, are in Figure 2(left). The BOW+POS hybrid setup is only applicable in Comrafs.

dataset is comparable with the number of POS trigrams, so in analogy to the previous model, we perform a 2-way clustering with the Comraf graph of Figure 1(c) and the objective $I(\tilde{D};\tilde{W})$.

4. BOW+POS hybrid: We combine contextual information of BOW and stylistic information of POS n-grams into a 3-way clustering model, where we simultaneously cluster documents, words and bigrams of POS tags. Over the Comraf graph of Figure 1(d), we maximize the sum $I(\tilde{D};\tilde{S}) + I(\tilde{D};\tilde{W})$.

5.1 Dataset

We evaluate our models on the British National Corpus (BNC) [Burnard, 2000]. We employ David Lee's ontology of BNC genres [Lee, 2001] with 46 genres covering most aspects of modern literature such as *fiction prose, biography, technical report, news script* and others. To perform fair evaluation using clustering accuracy (Section 4), we choose 21 largest categories, for each of which we uniformly at random choose 32 documents, so our resulting dataset consists of 672 documents. The BNC texts are represented in SGML. We remove all markup, lowercase the text, and delete stopwords and low frequency words. All words in the BNC corpus are semi-manually tagged using 91 POS tags, four of which refer to punctuation. The resulting dataset has 63,634 unique words; and 5864 POS bigrams. Since the overall number of unique POS trigrams and fourgrams is prohibitively large, we apply more aggressive term filtering: we consider trigrams that appear in at least 10 documents (44,499 trigrams overall) and fourgrams that appear in between 10 and 99 documents (114,476 fourgrams).

5.2 Results

We compare the results of our clustering model with the results of k-means (Weka implementation), as well as of Latent Dirichlet Allocation (LDA)—a popular generative model for unsupervised learning. We use Xuerui Wang's LDA implementation [McCallum *et al.*, 2005] that performs Gibbs sampling with 10000 sampling iterations. Table 1 summarizes the results which appear to be surprisingly good for an unsupervised method, given that the result of a random assignment of documents into 21 clusters would be about 5% accuracy. As shown, the 3-way Comraf model significantly outperforms other (1-way) models. Figure 2 shows results of stability tests of 2-way Comraf models: (left) the POS n-gram setup; (right) the BOW setup. As shown on the left figure, the POS bigrams setup is preferable over the other POS tuples: it is more effective than unigrams, and almost as effective as trigrams and fourgrams, while being much more efficient.

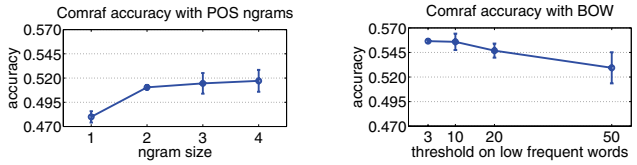

Figure 2: **Clustering by genre.** Comraf clustering accuracy as a function of: (left) size of POS n-gram (1-grams, 2-grams, 3-grams and 4-grams); (right) threshold on low frequency words—a point i on the X axis means that in this experiment words that appear in less than i documents are removed.

6 Clustering by sentiment

In clustering by authors' sentiment, data categories correspond to different levels of the authors' attitude to the discussed topic (e.g. liked/disliked, satisfied/unsatisfied etc.). The categories can be finer grained (strongly liked / somewhat liked etc.)—as long as it is possible to distinguish between two adjacent categories.

Following the procedure described in Section 2, after choosing the number of clusters and particular feature types, the user is asked to select a few seed features for each category. For clustering by sentiment, as well as for close tasks of clustering by authors' mood or by familiarity with the topic, relevant feature types may be words or word n-grams (i.e. semantic features). However, for other quite similar tasks, e.g. clustering by authors' age, not only semantics but also syntax can matter: children, for instance, use certain words more often than adults do; children also tend to use primitive (and sometimes erroneous) syntactic constructions ("me going bye-bye" etc.). In this paper, for simplicity, we experiment with word features only.

The task of selecting seed words has two issues. First, it is easier to come up with words that correspond to *extreme* sentimental categories ('spectacular', 'horrible'), but it is difficult to choose seed words for intermediate, mild categories. Nevertheless, as we will see in Section 6.2 users usually succeed in accomplishing this task. Second, in our early experiments, users consistently tended to choose words that were out of the vocabulary of a given dataset. Inspired by Liu *et al.* [2004], we decided to provide the users with a word list, to narrow her search only to the dataset vocabulary. Unlike Liu *et al.* [2004], whose task is topical clustering, we cannot automatically predict which words would be relevant. Instead, we employ Zipf's law and provide the user with a list of words from the interior of the frequency spectrum. We anticipate such a list to contain the most relevant seed words.

We then perform an iterative process of clustering that allows user's involvement in between clustering iterations. We apply a 2-way Comraf model (see Figure 1c): we first cluster documents that contain the selected seed words and then we cluster all words of these documents. In the latter step, our seed word groups are enriched with new words that have been clustered together with the original seed words. The user is

then asked to edit the new seed word groups, in order to correct possible mistakes made by the system (word removal, relocation and addition is allowed). By this, a clustering iteration is completed and the next iteration can be executed.

Since the seed word groups have been enlarged, we can expect that a set of documents that contain these seed words is now larger as well, so that the clustering process will cover more and more documents from iteration to iteration. The process stops when no more documents are added to the pool. Documents that have never been covered (the ones that contain no seed words from the largest seed word groups) are considered to be clustered incorrectly. An alternative approach to guarantee the algorithm's convergence would be to require enlargement of seed word groups such that at least one document is added to the clustering at each iteration. The algorithm would then stop when the entire dataset is covered. We choose the former approach because (a) we do not want to put additional constraints either on the user or on the Comraf clustering model; (b) in each real-world dataset there can be documents whose sentimental flavor is hard to identify – it would not be beneficial to force such documents into any of the sentimental clusters.

6.1 Experimental setup

We evaluate our interactive clustering system on a dataset of movie reviews. Our dataset consists of 1613 reviews written on *"Harry Potter and the Goblet of Fire (2005)"* that we downloaded from `IMDB.com` in May 2006.[1] The data was preprocessed exactly as the BNC corpus. We ignore reviews that do not have rating scores assigned by the user. The IMDB's scoring system is from 1 (the worst) to 10 (the best). Based on our extensive experience with `IMDB.com`, we translate these scores into four categories as follows: scores 1 to 4 are translated into the category *strongly disliked* (292 documents), scores 5 to 7 are translated into *somewhat disliked* (454 documents), scores 8 and 9 into *somewhat liked* (447 documents), and score 10 is translated into the category *strongly liked* (420 documents). We do not introduce a neutral category because there are very few neutral reviews on `IMDB.com`.

On the task of clustering by sentiment, we compare our method's performance with that of an SVM classifier trained on 22,476 movie reviews. The training data for the SVM consisted of reviews of 46 popular Hollywood movies released in 2005, of the same genre as *Harry Potter*. The reviews and genre labels of movies are obtained from `IMDB.com`. Again, we ignore reviews without user-assigned rating.

The system is evaluated on five users who are familiar with the task of document clustering. The users were explained the idea behind interactive clustering and provided a brief description of the dataset. They were given a list of 563 words that appeared in $50 \leq n < 500$ documents in our dataset. The users proceeded as described in Section 2. Also, we construct an *oracle* as follows: for each category t we select 25 most frequent words that belong to a given list of sentimental words[2] and their distribution over the categories has a peak at t. Unlike human users, the oracle does not provide feedback between clustering iterations. To some extent, the oracle's performance can be considered as an upper bound to results obtained in practice, when a human user is involved.

We perform a *simulated classification* (SC) experiment analogous to the one of Liu *et al.* [2004] (see a description in Section 1), where the seed words are provided by our oracle. We replace an ad-hoc kNN-like clustering in Liu *et al.*'s implementation by our effective Comraf clustering, and a Naive Bayes classifier by an SVM.

Doc repres.	k-means	LDA	Comraf	SVM
BOW	28.2	37.0 ± 0.2	40.3 ± 0.8	39.1 ± 0.3
Sentim. list	29.0	40.2 ± 0.5	43.0 ± 0.9	41.3 ± 0.6
Interactive clustering (Oracle)			$\mathbf{47.1 \pm 0.2}$	n/a
Simulated classification (Oracle)			46.3 ± 0.1	

Table 2: **Clustering by sentiment.** Clustering accuracy vs. classification accuracy. Standard error of the mean is shown after the $\pm$ sign.

6.2 Results

Table 2 summarizes our observations. Surprisingly, with BOW features, our Comraf clustering method performs as well as an SVM trained on a large amount of data (Row 1). A good performance of our unsupervised method (with BOW) indicates that the constructed topical clustering sheds some light on reviewers' sentiments, which can occur when the reviewers have a consensus on certain aspects of the movie, e.g. liked the actors but disliked the plot etc.

After feature selection according to our list of sentimental words, the Comraf achieves a significant boost in accuracy surpassing the SVM (Row 2). Using an oracle in our interactive clustering setup (Row 3) improves the performance even further, while the SC result (Row 4) is only slightly (but significantly) inferior. These two results are close because the training set of SC is identical to the clustering constructed at the first iteration of the Comraf algorithm. As its size appears to be over 3/4 of the entire dataset, there is almost no room for the actual diversity in performance of the two methods.

Figure 3 (left) shows the accuracy (micro-averaged over the classes) for each user and each iteration. For three of the five users, selection of the initial seed words is sufficient to obtain significantly higher accuracy than the best result of the SVM. User 2 has significantly lower accuracy than the baseline to begin with, but over the two correction steps is able to provide the necessary feedback so as to obtain an improvement in accuracy, equalling the baseline. We found that User 2 was fairly conservative in her assessment of terms in the

[1] Bo Pang [Pang and Lee, 2005] maintains a popular dataset of movie reviews that, unfortunately, does not fully correspond to our task because (a) we want to differentiate the problem of clustering by sentiment from the topical clustering—for this reason our dataset contains reviews written on *one* movie only, so that the *topic* of all the reviews is potentially the same; (b) movie ratings in Bo Pang's dataset are extracted from the reviews' text, which is an error-prone procedure, whereas in our dataset the ratings are assigned by the reviewers using an HTML form which leaves no room for errors.

[2] Our list of 4295 sentimental words was obtained as described in [Eguchi and Lavrenko, 2006].

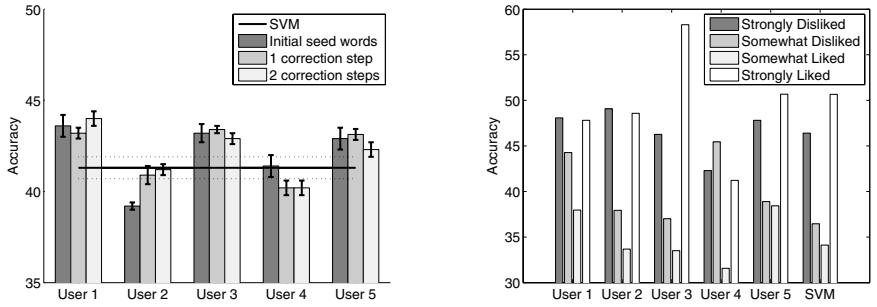

Figure 3: **Interactive clustering by sentiment.** Clustering accuracy over various users: (left) over interactive learning iterations (with original seed words only, after one correction step and after two correction steps). The horizontal line is SVM performance (after feature extraction using a given list of sentimental words, and after training on over 20K documents); (right) over categories of the dataset after two correction steps.

beginning marking only 26 terms, while User 1 (the one with the best average performance) marked 58 terms, 23 of which were in common with User 2. User 4 reported that she aggressively removed words at the first correction step, which caused a noticeable drop in the performance.

Figure 3 (right) shows the accuracy per class, per user at the end of 3 iterations. User 1 and User 2 have near identical accuracies on the two extreme categories (*strongly liked* and *strongly disliked*), but User 1 has higher accuracies on the intermediate categories, resulting in higher micro-averaged accuracy. It is apparent from this figure that users are able to come up with good features for the two extreme categories, but have difficulty with the intermediate categories. The figure also shows the performance of SVM (with sentiment features). It is interesting to note that the SVM's pattern of behavior is almost identical to the interactive Comraf's.

7 Conclusion

We have introduced an interactive clustering method that allows on-the-fly clustering of text collections according to any (especially non-topical) criterion the user can come up with. In our method, the user's prior knowledge on the importance of features is incorporated into the multi-modal clustering. We apply our method to clustering movie reviews by sentiment. It takes the user less than 15 minutes to choose initial seed words using which our system significantly outperforms an SVM trained on a large amount of data. The subsequent correction steps however are often unnecessary. We also test our system on clustering by genre where seed features cannot be chosen. Instead, we cluster documents together with their contextual and stylistic features and achieve good results.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, and in part by the Ministry of Education, Culture, Sports, Science and Technology of Japan under grant number KAKENHI-17680011.

References

[Basu *et al.*, 2004] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *SIGKDD-10*, pages 59–68, 2004.

[Bekkerman and Sahami, 2006] R. Bekkerman and M. Sahami. Semi-supervised clustering using combinatorial MRFs. In *ICML-23 Workshop on Learning in Structured Output Spaces*, 2006.

[Bekkerman *et al.*, 2005] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way distributional clustering via pairwise interactions. In *ICML-22*, pages 41–48, 2005.

[Besag, 1986] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3), 1986.

[Burnard, 2000] L. Burnard. User reference guide for the British National Corpus. Technical report, Oxford University Computing Services, 2000.

[Dayanik *et al.*, 2006] A. Dayanik, D. D. Lewis, D. Madigan, V. Menkov, and A. Genkin. Constructing informative prior distributions from domain knowledge in text classification. In *SIGIR-29*, pages 493–500, 2006.

[Eguchi and Lavrenko, 2006] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *EMNLP*, 2006.

[Lee, 2001] D. Y. W. Lee. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3), 2001.

[Liu *et al.*, 2004] B. Liu, X. Li, W. S. Lee, and P. S. Yu. Text classification by labeling words. In *AAAI-19*, pages 425–430, 2004.

[McCallum *et al.*, 2005] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. In *IJCAI-19*, pages 786–791, 2005.

[Pang and Lee, 2005] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL-43*, pages 115–124, 2005.

[Raghavan *et al.*, 2005] H. Raghavan, O. Madani, and R. Jones. InterActive feature selection. In *IJCAI-19*, pages 841–846, 2005.

[Turney, 2002] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL-40*, pages 417–424, 2002.

[Willett, 1988] P. Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5):577–597, 1988.