An Exploration into the Use of Contextual Document Clustering for Cluster Sentiment Analysis

Niall Rooney, Hui Wang
University of Ulster
{nf.rooney,h.wang}
@ulster.ac.uk

Fiona Browne
Queen's University,Belfast
f.browne@qub.ac.uk

Fergal Monaghan, Jann Müller, Alan Sergeant, Zhiwei Lin, Philip Taylor
SAP Research Belfast
{fergal.monaghan,
jann.mueller,
alan.sergeant,
zhiwei.lin,
philip.taylor}@sap.com

Vladimir Dobrynin
St Petersburg State University
v.dobrynin@bk.ru

Abstract

In this paper we consider whether the thematic document clustering approach of Contextual Document Clustering is able to capture the overall sentiment of a cluster of documents. We provide a novel mechanism to determine the sentiment of a cluster based on the latter approach and assess the approach on three data sets formed from the NY Times annotated corpus. We demonstrate that CDC does provide a strong tendency to capture the sentiment of a cluster.

1 Introduction

Sentiment analysis or opinion mining is a recent area of text classification research which tries to determine the opinion that a section of text expresses. Esuli & Sebastiani (2005) describes three subtasks:

- determining whether a given piece of text has a factual nature, neutral nature or whether it expresses an opinion on its material (the Subjective-Objective (SO) polarity of the text)
- determining whether a given piece of text expresses a positive or negative opinion on its subject matter orientation (the Positive-Negative (PN) polarity of the text)
- determining the strength of the subject matter orientation

Turney & Littman (2003) make no distinction between the latter two sub-tasks and propose a measure of semantic orientation which indicate both the direction and intensity of a text. To capture this, they focus on the semantic orientation of a word which they capture by measuring the strength of association with a set of seed words (with either absolute positive or negative polarity). They propose two measures for strength of association based on point-wise mutual information and latent semantic analysis estimated from given corpora. Other mechanisms for semantic orientation have focused on the linguistic constraints on the orientation of adjectives (e.g. the word "and" usually conjoins adjectives of the same orientation) (Hatzivassiloglou & McKeown, 1997) or that synonymous words have similar orientation

Proceedings of Recent Advances in Natural Language Processing, pages 140–145,
Hissar, Bulgaria, 12-14 September 2011.

(Esuli & Sebastiani, 2005). The aforementioned mechanisms try to give an absolute value for the orientation of a word regardless of its context of use. Wilson et al.(2009) present a two stage classification approach to determine the contextual polarity of subjective clues (words which have been part of annotated subjective expressions) in a corpus. They based this on features primarily as a consequent of local dependency relationships (parent-child) in sentences (although they do use other features mainly at a sentence level). More recent directions in a sentiment analysis for text classification have focussed on the use of unsupervised modelling approaches for text classification. Much of this work has focussed on extending topic modelling approaches such as Probabilistic Latent Semantic Indexing (Mei et al., 2007) or Latent Dirchlet Allocation (Lin & He, 2009) to incorporate the use of sentiment as a variable.

To our knowledge, little work has focussed on determining the sentiment of a cluster rather than the individual documents. Dobrynin et al.(2004,2006,2008) proposed the unsupervised mechanism of Contextual Document Clustering (CDC) that by discovering distinct and relevant contexts, allows for the hard partitioning of documents in a corpus into theme based clusters. A "theme" is an implicit concept and can be considered as equivalent in intent to its lexical definition. CDC considers words in a corpus as any character sequence occurring between separators (either whitespace or punctuation marks) in any text in document. A term in a document is a constrained character sequence based on a regular expression, so in general the set of terms is a subset of the set of words. Also a word cannot be a stop word. A context is a probability distribution of co-occurring terms in documents given a context term. CDC's partitioning of documents, is based on information theoretic considerations of semantic similarity between a document and a context. There exists a logarithmic relationship between a context term's document frequency and the context's entropy. As such, the final choice of context terms and their respective contexts are based on the grouping of context words into a fixed number N_{dfg} of document frequency intervals, and the entropy of their associated context. Contexts are chosen from each interval in a round robin fashion in order of least entropy from each group. In total N_c are chosen. To allow for the fact that after this step, certain contexts may still be too similar based on a comparison of their distributions, merging steps are carried out to merge similar contexts.

In this paper we assess whether CDC by capturing theme related documents within a given cluster, also intrinsically captures the theme's sentiment and would allow for a categorization of a cluster based on sentiment. We hypothesis that if this is the case, an independent measure of a cluster's sentiment will show a high likelihood that a cluster to be either positive or negative in sentiment overall or be a mixture of positive and negative sentiments so that the overall sentiment is neutral. In the latter case this would allow for a further decomposition of sentiment analysis based on sub-regions of the cluster. In the small minority of cases will a cluster be composed solely as a mixture of neutral sentiments. In general, all clusters will contain a mixture of negative and positive sentiments, so we are assessing if the sum polarity tends to be mainly positive or negative i.e. a majority of clusters will either be positive or negative in sentiment.

2 Methodology

For each cluster formed by CDC, it is possible to derive a set of base concepts that provide tag descriptors of the cluster. These tags provide a semantic description of the cluster. Our assumption is that these descriptors also form the basis for determining the overall sentiment of a cluster by the additional use of lexicons of known positive and negative words. This allows a simpler determination of the cluster's sentiment. If it can be shown that for a majority of clusters, a cluster has either a positive or negative sentiment, this provides support for the hypothesis given in section 1.

Each cluster C has a cluster description consisting of a set of cluster tags T and the cluster contains a set of D_c documents. Each document, $d \in D_c$ consists of a set S_d of sentences where a sentence is determined by known boundaries such as punctuation marks. A tag is a contiguous sequence of two or three word phrases. Let Pos be the set of known positive words. These are words that exist in the original lexicon of positive words and exist in the corpus. Let Neg be the set of known negative words. These are words that exist in the original lexicon of negative words and exist in the corpus. Let N_{dc} be the number of documents in cluster C . Let df_w be the number of documents in the cluster for which the word frequency of a word w within a document is non-zero.

$$df_w = \; |\{d \in D_c : tf(w,d) > 0\}|$$

Let df_t be the number of documents in the cluster for which the tag frequency $pf(t,d)$ of the tag t within the document is non-zero:

$$df_t = \; |\{d \in D_c : pf(t,d) > 0\}|$$

The document frequency of documents $df_{t \wedge w}$ which contain both a tag t and word w within the same sentence (in the same *vicinity*), is defined as:

$$df_{t \wedge w} = |\{d \in D_c : \exists s \in S_d : tf\ (w,s).pf\ (t,s) > 0\}|$$

The cluster sentiment CS is calculated as follows based on Pointwise Mutual Information (PMI) between a word $w \in Pos$ or a word $w \in Neg$ and a tag t summed over all tags:

$$CS = \sum\nolimits_{t \in T} TS(t)$$

$$TS(t) = \log_2\left(\frac{\prod_{w \in P} df_{t \wedge w}}{\prod_{w \in P} df_t . df_w}\right) - \log_2\left(\frac{\prod_{w \in N} df_{t \wedge w}}{\prod_{w \in N} df_t . df_w}\right)$$

In effect, cluster sentiment is the summation of the tag sentiments.

This formula is based on Turney & Littman study (2003) where we are replacing occurrences of a co-occurring word (with another word) with a co-occurring tag t. We only consider tags that do not contain either positive or negative words as part of their phrasal text.

We assume a cluster has positive sentiment if,

$$CS > Thres$$

neutral if,

$$Thres >= CS >= -Thres$$

and negative if,

$$CS < -Thres$$

Normally the threshold value is 0, however we allow an admittedly arbitrary greater value than 0 to indicate that weakly positive or negative cluster sentiment should be considered neutral. We refer to this calculation for clusters as **CS-standard**. A standard lexicon may also have a measure of the subjective strength of the word whether a word in most contexts is seen as strongly or weakly subjective.

To allow for this factor we modified $TS(t)$ to include a subjectivity factor for lexicon words, where words which are strongly subjective have a different factor to words that are weakly subjective.

$$TS(t) = \log_2\left(\frac{\prod_{w \in P} \alpha_w . df_{t \wedge w}}{\prod_{w \in P} df_t . df_w}\right) - \log_2\left(\frac{\prod_{w \in N} \alpha_w . df_{t \wedge w}}{\prod_{w \in N} df_t . df_w}\right)$$

We refer to this calculation for clusters as **CS-subj**. The factor, α_w was set to 2.0 for strongly subjective lexicon words and to 1.0 for weakly subjective.

CS-standard considers all tags to be of equal importance. Based on the tag document frequency within a cluster, it is possible to give each tag a weighting normalized by the tag frequency range:

$$CS = \sum\nolimits_{t \in T} \lambda_t TS(t)$$

$$\lambda_t = 0.5 + \left(0.5 * \frac{df_t - \min_{tag \in T} df_{tag}}{\max_{tag \in T} df_{tag} - \min_{tag \in T} df_{tag}}\right)$$

We refer to this mechanism as **CS-rank**. This approach gives a weighting for each tag between 0.5 and 1.0, so that tags with higher document frequency have greater weighting.

3 Evaluation

The choice of data set was determined by two factors. Firstly, the data set had to contain sufficient documents to form a set of information-rich contexts and hence clusters. Secondly, the nature of the data set has a high likelihood of expressing a mixture of subjective opinions. For this purpose, we chose data from the NY Times annotated corpus (Sandhaus, 2008). We considered 3 subsets of data for the respective years of 2005 (Nyt-2005), 2006 (Nyt-2006) and 2007 (Nyt-2007) and ran the same evaluation for each corpus. We based each evaluation on the subjectivity lexicon provided by Wilson et al. (2005) which lists a set of words with either positive or negative polarity and a measure of subjective strength (either strong or weak). This latter feature was the basis for the setting of α_w in CS-subj. In total there are 2304 positive words and 4145 negative words. Not all words were present in each of the 3 corpora and such words were ignored. Table 1 summarizes the data characteristics for the three data sets and indicates that the parameters are stable for each evaluation, not surprisingly as there is no variation in the nature of the data. Nyt-2007 has fewer documents as data was only recorded up to April, 2007. The *Thres* value was set to 5.0 indicating that clusters with sentiment only weakly positive or weakly negative, we considered as neutral.

Data set	**Number of documents**	**Number of clusters**	**Number of positive words in corpus**	**Number of negative words in corpus**
Nyt-2005	89975	1363	2172	3796
Nyt-2006	87029	1339	2165	3785
Nyt-2007	39950	1396	2132	3675

Table 1 Data set characteristics

There appears to be an imbalance between the number of positive and negative words but this imbalance is less pronounced if we consider only words in the lexicon that occur in the vicinity of cluster tags (only such words contribute to the evaluation scores). This is shown in Table 2.

Data set	**Number of positive words in vicinity of a given tag**	**Number of negative words in vicinity of a given tag**
Nyt-2005	1790	2657
Nyt-2006	1912	3010
Nyt-2007	1912	3095

Table 2 Lexicon words used in Evaluations

As described in the Introduction, CDC requires an apriori setting of how many distinct contexts to select N_c and the number of document frequency intervals N_{dfg} (Rooney *et al.*, 2006). Note that there can be fewer contexts formed than requested due to merging of similar contexts and fewer clusters also due to non-assignment of documents to given contexts. In each evaluation, N_c was set to 2000 and N_{dfg} to 7, as previous work has shown these values to be appropriate settings for these sizes of data sets. We then calculate the number of positive, negative or neutral clusters and express the relative number of clusters as percentages. This process was carried out for each evaluation and results were averaged over the 3 years. The average of the evaluations is shown in Figure 1.

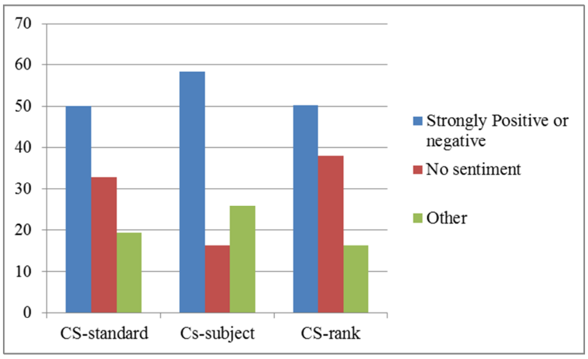

Figure 1: Cluster sentiment averaged over nyt_2005, nyt_2006,nyt_2007

Clearly there is a majority of clusters that are either positive or negative in sentiment in the case of both CS-standard and CS-subj so that our hypothesis is justified. There is very little to distinguish these two mechanisms with CS-subj returning a slightly elevated percentage of positive clusters and a similarly decreased percentage of negative sentiment clusters and this was reflected not only in the averages but in the individual evaluations. CS-rank shown a somewhat different profile with there still been a majority of clusters being identified as positive or negative, but a relative reduction in the percentage of positive clusters and a relative increases in the percentage of negative clusters. However investigation into each data set showed the consistent pattern of increasing the number of neutral clusters and we can consider this mechanism of ‘smoothing’ the individual contribution of each tag.

Further evidence is provided for our hypothesis when we examined clusters deemed as neutral. We consider each neutral cluster as belonging to one of two categories: *no sentiment* if in fact the overall sentiment is 0 which only happens if no sentiment value is calculated for given cluster tags and *sentiment* otherwise. Figure 2 shows the results of this categorization average over the 3 evaluations.

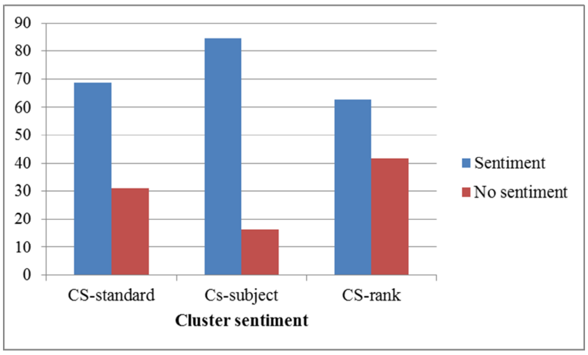

Figure 2: Neutral Cluster decomposition averages averaged over nyt_2005, nyt_2006, nyt_2007

Regardless of the cluster sentiment measure, only a minority of neutral clusters are truly neutral and show no tag sentiment. Of course in the other case the cluster would need to be decomposed into smaller regions to allow for the discovery of regions of either positive or negative sentiment, if we are not to regard the cluster as "neutral". CDC provides graph based mechanisms to structure the content of clusters whose use for sentiment analysis will be explored in future work.

It is not uncommon for CDC to form clusters based on themes which share tags, as tags are not a description of the intrinsic theme or context, but simply indicators of the cluster's content. As this is the case, it was of interest to consider whether clusters that have a high degree of similarity in their tags could have different cluster sentiment classification. We considered a pair of clusters that shared at least 70% of tags relative to the first cluster in a pairing as highly similar. Table 3 shows a summary of the outcome.

Clearly there is evidence that between 23 to 29 percent of cluster pairings have different sentiment, again highlighting the use of cluster tags as the basis for determining cluster sentiment. The tags by themselves do not give any indication of the overall cluster sentiment, but individually they are the basis for determining tag sentiment as contributors to overall cluster sentiment.

Data set	**Number of highly similar clusters**	**Number of highly similar clusters with different cluster sentiment (Percentage)**
Nyt-2005	449	130 (29%)
Nyt-2006	473	136 (29%)
Nyt-2007	336	78 (23%)

Table 3 Similar pairs of clusters and Number with differing cluster sentiment

We do not have an independent means of assessing the strength of our approach to tag sentiment and hence cluster sentiment - we would need human assessors to provide a qualitative evaluation, but we have seen a considerable number of examples whereby the tag sentiment for different clusters is clearly reflected the documents that contain these tags. By way of example, consider the following two highly similar clusters <13820,15095> drawn from the Nyt_2006 evaluation, where the clusters identifiers are as a result of the CDC process. The following table shows the tagging for cluster 13820.

Tag list for cluster
tom glavine
orlando hernández
omar minaya
pedro martínez
dominican republic
carlos delgado
willie randolph
shea stadium

Table 4 Cluster tags for cluster 13820

Clearly the cluster is topically related to "baseball". The tag list is much longer for 15095 with 7 of the tags from 13820, also occurring for 15095. 15095 is also topically related "baseball" – how they vary thematically is intrinsic to the context, which is hard for us to convey as they are probabilty distribution in words but clearly the themes have some level of similarity. If we consider the tag "pedro martínez", this has tag sentiment -15.78 in 15095 and 39.87 in 13820. The given tag occurs in 4 documents in 15095 and 2 in 13820. Table 4 shows the titles for these documents (the content of a document is a concatenation of both its title and its body of text) which demonstrates that the tag “pedro martínez” has a strong difference in sentiment for these two clusters, allowing for the fact that the judgment is based on titles only.

Cluster: 15095	Cluster: 13820
Randolph Lets Bygones Be Bygones Martínez May Have to Consider Retiring Martínez Takes It Step by Step, Gingerly Martínez on Hill, But Not in Shape	No News on Martínez, and Mets Say That's Good Martínez: Good Guy In Mets' Black Hat

Table 5 Document titles containing the same tag “pedro martínez” but different clusters

4 Conclusions

We have shown in this paper that for the given type of data, CDC is likely to form clusters reflecting an intrinsic polarity in sentiment. This may only be reflected in news articles where the expression of opinions is commonplace and we propose considering other data sets of a less opinionated nature to see how they compare. In future work, we aim to benchmark our approach against other approaches to document clustering to see if CDC is superior in this aspect.

References

Dobrynin, V. Patterson, D. Rooney, N. (2004): Contextual Document Clustering. ECIR 2004: 167-180

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of words through gloss analysis. *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management,* 617–624.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics,* 174-181.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. *Proceeding of the 18th ACM Conference on Information and Knowledge Management,* 375-384.

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. X. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. *Proceedings of the 16th International Conference on World Wide Web,* 171-180.

Rooney, N., Patterson, D., Galushka, M., & Dobrynin, V. (2006). A scaleable document clustering approach for large document corpora. *Information Processing & Management, 42*(5), 1163-1175.

Rooney, N., Patterson, D., Galushka, M., Dobrynin, V., & Smirnova, E. (2008). An investigation into the stability of contextual document clustering. *Journal of the American Society for Information Science and Technology, 59*(2), 256-266.

Sandhaus, E (2008) *The New York Times Annotated Corpus*. Linguistic Data Consortium, Philadelphia.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS),* 315-346.

Wilson, T., Wiebe & Paul Hoffmann, P.(2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of HLT/EMNLP 2005,Vancouver, Canada.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics, 35*(3), 399-433.