When to Generalise - A Case-Based Approach to Text Modelling

Sadiq Sani, Nirmalie Wiratunga, Stewart Massie, and Robert Lothian

School of Computing,
Robert Gordon University,
Aberdeen AB25 1HG, Scotland, UK
{s.a.sani,n.wiratunga,s.massie,r.m.lothian}@rgu.ac.uk

Abstract. The variation in natural language vocabulary remains a challenge for textual case representation as the same idea can be expressed in many different ways. Thus document representations typically rely on generalisation to map low-level lexical expressions to higher level concepts in order to capture the inherent semantics of the documents. Term-relatedness measures are often used to generalise document representations by capturing semantic relationships between terms. However, generalisation is not guaranteed to improve retrieval performance all the time. Extracting pairwise term-relatedness is expensive therefore, it is important to be able to predict, given a dataset, whether or not applying generalisation will be beneficial. If not, we can avoid the overhead of having to extract term relatedness values in the first place. In this work we present a CBR approach that predicts, given a text classification dataset, whether or not using generalisation will improve text classification performance. Our evaluation shows that our CBR approach outperforms a ZeroR baseline that predicts always using generalisation.

Keywords: Meta Case-Based Reasoning, Textual Case-Based Reasoning, Generalisation, Term Relatedness

1 Introduction

A solution commonly used to address the problem of variation in natural language vocabulary during text retrieval is to introduce measures of semantic relatedness between terms. This provides a mapping between different lexical expressions of a similar idea into conceptual groups that capture the inherent meanings of documents. The result is the generalisation of document representations away from low-level expressions to high-level semantic concepts. Different approaches have been proposed for obtaining term relatedness knowledge. These range from using knowledge rich (extrospective) sources (e.g. lexical databases, Wikipedia and the World Wide Web) to knowledge light (introspective) techniques that use statistics of term co-occurrences in a corpus. Despite their simplicity, statistical techniques have so far provided the best performance in text retrieval evaluations [2]. One reason for this is that corpus co-occurrence is particularly helpful

for estimating domain specific relationships between terms [3]. Also, statistical approaches are able to capture relationship types other than similarity e.g. the association between 'bank' and 'money'.

Although generalisation has proven quite useful, it remains to be determined whether generalisation always improves retrieval performance. Thus, the aim of this work is to address two important questions:

1. Does generalisation always work?
2. If no, can we predict when it is likely to work?

We address the first question by investigating the performance of co-occurrence based generalisation on a number of text classification datasets. To address the second question, we investigate several attributes of text classification datasets that are predictive of the performance of generalisation. We then use these attributes in a case-based system to predict, given a text classification dataset, whether or not to apply generalisation. Being able to accurately predict when generalisation is not likely to improve performance means that we can conveniently avoid the overhead of having to extract term relatedness knowledge in the first place.

The rest of this paper is structured as follows, in Section 2 we describe popular statistical approaches for extracting term relatedness. Section 3 describes the datasets we used in our evaluation. In Section 4 we present an evaluation of the performance of term relatedness on text classification. In Section 5 we present several metrics for capturing different attributes of our datasets and in Section 6 we present our CBR approach for predicting when to use term relatedness using these attributes. We present related work in Section 7. Conclusions are presented in Section 8.

2 Term Relatedness From Corpus Co-occurrence

The general idea of introspective approaches is that co-occurrence patterns of terms in a corpus can be used to infer semantic relatedness such that the more two terms occur together in a specified context, the more related they are. In the following sections, we describe three different approaches for estimating term relatedness from corpus co-occurrence.

2.1 Document Co-occurrence

Documents are considered to be similar in the vector space model (VSM) if they contain a similar set of terms. In the same way, terms can also be considered similar if they appear in a similar set of documents. Given a standard term-document matrix D where column vectors represent documents and the row vectors represent terms, the similarity between two terms can be determined by finding the distance between their vector representations. The relatedness between two terms, t_1 and t_2 using the cosine similarity metric is given in equation 1.

$$CoocSim(t_1, t_2) = \frac{\sum_{i=0}^{n} t_{1,i} t_{2,i}}{|t_1||t_2|} \tag{1}$$

2.2 Latent Semantic Indexing

Latent Semantic Indexing (LSI) is a technique that uses singular-value decomposition (SVD) to exploit co-occurrence patterns of terms and documents to create a semantic concept space which reflects the major associative patterns in the corpus [7]. In this way, LSI brings out the underlying latent semantic structure in texts.

Given a term-document matrix D, SVD is used to decompose D into three matrices: U, a term by dimension matrix; S a diagonal matrix of singular values; and V, a document by dimension matrix. This decomposition is shown in equation 2. The number of dimensions n is the rank of the original term-document matrix D.

$$D = U \times S \times V \tag{2}$$

The U, S, V matrices are truncated to k dimensions which represent the k most important concepts in the term-document space. A new term-document representation generalised to this concept space can then be obtained by multiplying the rank-reduced U, S, V matrices using equation 2.

2.3 Normalised Positive Pointwise Mutual Information

The use of mutual information to model term associations is demonstrated in [4]. Given two terms t_1 and t_2, mutual information compares the probability of observing t_1 and t_2 together with the probability of observing them independently as shown in equation 3

$$PMI(t_1, t_2) = log_2 \frac{P(t_1, t_2)}{P(t_1)P(t_2)} \tag{3}$$

If a significant association exists between between t_1 and t_2, then the joint probability $P(t_1, t_2)$ will be much larger than the independent probabilities $P(t_1)$ and $P(t_2)$ and thus, $PMI(t_1, t_2)$ be greater than 0. Positive PMI is obtained by setting all negative PMI values to 0. The probability of a term t in any context can be estimated by the frequency of occurrence of t in that context normalised by the frequency of all words in all contexts.

$$P(t) = \frac{f(t)}{\sum_{j=1}^{N} \sum_{i=1}^{N} f(t_i, t_j)} \tag{4}$$

PMI values do not lie within the range 0 to 1. Thus we need to introduce a normalisation operation. We normalise PMI as shown in equation 5.

$$NPPMI(t_1, t_2) = \frac{PPMI(t_1, t_2)}{-log_2 P(t_1, t_2)} \tag{5}$$

3 Datasets

Datasets used for evaluation were mostly obtained from standard text classification corpora e.g. 20 Newsgroups, Ohsumed, Reuters V1 and Movie Reviews. Additional classification datasets were created from incident reports crawled from the overnment of Western Australia's Department of Mines and Petroleum website. These corpora are described in detail in the following paragraphs.

20 Newsgroups corpus is a collection of 20,000 documents collected from Newsnet newsgroups messages. The collection is partitioned almost equally into 20 classes of 1,000 documents each, according to newsgroup topics. For example, the class sci.space contains messages relating to space.

Ohsumed is a subset of MEDLINE, an online database of medical literature, and comprises a collection of 348,566 medical references from medical journals covering a period from 1987 to 1991. The Ohsumed collection is unequally divided into 23 classes according to different disease types e.g. Virus Diseases. The classification of documents in this collection is non-disjoint which means the same document can be categorised under two or more different classes if it is relevant to all those classes. For our experiments, we selected only documents that belong to a single class.

Reuters Volume 1 corpus is an archive of 806,791 news stories provided for research purposes by the global news provider, Reuters. The collection comprises all news stories produced by Reuters journalists within a one year period starting from August, 1996. Documents within the collection are tagged with descriptive metadata specifying codes for topic, region and industry sector. Topic codes represent the subject area of each news story. Industry codes are used to indicate the type of business or industry referred to by the news story. Region codes indicate the geographical region referred to in the news story. Only topic codes and industry codes where used when creating datasets for our evaluation.

Moview Reviews is a sentiment classification corpus containing 1400 reviews of movies from the Internet Movie Database (IMDB). About half of these reviews are classified as expressing positive sentiment while the other half is classified as negative. Accordingly, the classification task for this dataset is to determine the sentiment orientation of any given review.

Incident Reports corpus was created using incident reports crawled from the Government of Western Australia's Department of Mines and Petroleum website [1]. The corpus comprises reports describing several incident types e.g. fire related incidents and truck collisions.

[1] http://dmp.wa.gov.au

4 Measuring Performance of Generalisation

The aim of this evaluation is to determine the general performance of generalisation on a variety of text classification datasets. Particularly, we are interested in determining if generalisation consistently improves text classification accuracy.

4.1 Experiment Setup

Standard preprocessing operations i.e. stemming and stopwords removal were applied to our datasets. Chi^2 feature selection metric is used to limit our term-document space to the top 300 most informative terms for each dataset.

We compare the following algorithms: **BASE**, baseline BOW approach without term relatedness; **COOC**, term relatedness estimated from document co-occurrence (see Section 2.1); **NPMI**, term relatedness calculated using Normalised Positive Pointwise Mutual Information (see Section 2.3); **LSI**, term relatedness estimated from latent semantic analysis (see Section 2.2).

For text representation we use the framework we introduced in [13]. We report classification accuracy using a weighted kNN approach (with k=3) and using the cosine similarity metric to identify the neighbourhood.

4.2 Results

Classification results are shown in table 1. Values with the $^+$ sign correspond to a 1% or greater improvemnt in text classification accuracy compared to the baseline. Values with $^-$ on the other hand represent a 1% or greater decline in classification accuracy. The average difference between Cooc and Base is 1.26%, between Npmi and Base is 1.09%, and between Lsi and Base is 0.87%.

Although generalisation has resulted in improvement in most datasets (58% of the datasets using Cooc, 52% using Lsi and 51% using Npmi), it has remained neutral on many other datasets and even led to a decline in accuracy in others. Considering the additional cost of acquiring term-relatedness, it is important to empirically determine when it is beneficial to use term relatedness in text classification. A logical place to begin is by trying to find attributes of the datasets that correlate well with the difference in classfication performance after generalisation. Given these dataset attributes, it is then possible to build a KNN classifier to predict whether or not to use generalisation for a particular dataset. In the next section, we discuss several attributes of classfication datasets.

5 Dataset Attributes

In this section, we present various attributes that capture different characteristics of text classification datasets. We aim to use these attributes as a set of features that can be used with a KNN classifier for predicting when to use term relatedness in text classification. These attributes range from simple measures of average terms per document to more complicated measures of classification

Table 1. Classification accuracy of different generalisation techniques.

Dataset	Base	Cooc	Npmi	Lsi
Hardw	89.7	91.2^{+}	91.2^{+}	91.0^{+}
MedSp	95.3	93.0^{-}	95.6	90.4^{-}
CryptE	95.4	89.9^{-}	92.1^{-}	90.4^{-}
ChrisM	88.0	90.1^{+}	89.1^{+}	90.8^{+}
MeastM	94.1	93.8	94.1	94
GunsM	93.1	93.6	93.7	93.8
AutoC	94.0	94.9	96.2^{+}	94.8
BaseH	95.6	95.7	96.6^{+}	95.5
StratM	88.2	89.5^{+}	83.8^{-}	90.2^{+}
EntTour	94.6	95.6^{+}	95.3	95.3
EqtyB	95.7	95.5	95.0	95.5
FundA	90.7	92.1^{+}	90.6	92.3^{+}
InRelD	92.1	93.8^{+}	91.4	93.8^{+}
NProdRes	85.0	86.6^{+}	80.1^{-}	86.3^{+}
ProdNP	86.8	89.0^{+}	88.2^{+}	88.5^{+}
MarketA	89.0	88.8	89.2	88.9
MoneyC	94.6	94.2	93.3^{-}	94.2
OilGas	85.8	85.8	84.9	85.7
ElectG	87.8	84.1^{-}	83.2^{-}	83.7^{-}
FinI	86.0	86.8	84.5^{-}	86.7
MovieRev	71.2	78.9^{+}	81.4^{+}	68.3^{-}

Dataset	Base	Cooc	Npmi	Lsi
BactV	84.4	89.5^{+}	89.8^{+}	88.1^{+}
NervI	91.0	91.1	93.1^{+}	90.3
CardR	90.0	93.2^{+}	94.1^{+}	92.2^{+}
MouthJ	89.8	92.0^{+}	92.8^{+}	91.9^{+}
NeopE	92.1	94.3^{+}	93.8^{+}	94.0^{+}
DigNut	86.8	91.6^{+}	93.5^{+}	91.8^{+}
MuscS	82.8	87.3^{+}	91.2^{+}	86.5^{+}
EndoH	90.8	95.7^{+}	96.3^{+}	95.1^{+}
MaleF	92.4	95.0^{+}	95.3^{+}	95.4^{+}
PregN	90.3	89.1^{-}	91.4^{+}	88.9^{-}
ImmunoV	78.6	81.0^{+}	84.1^{+}	81.5^{+}
NervM	84.2	87.4^{+}	91.1^{+}	87.3^{+}
RespENT	86.6	87.8^{+}	91.3^{+}	88.5^{+}
SkinN	86.7	87.3^{+}	91.0^{+}	86.6
EndoNut	75.2	81.4^{+}	82.4^{+}	81.3^{+}
Fire	84.5	88.0^{+}	86.5^{+}	87.5^{+}
Collision	82.5	83.0	77.0^{-}	82.0
Rollover	78.0	76.5^{-}	76.0^{-}	76.0^{-}
CollRoll	85.5	81.0^{-}	81.5^{-}	84.0^{-}
MiscInc	83.0	84.0^{+}	80.5^{-}	83.0
CraneFP	84.5	87.0^{+}	82.0^{-}	87.5^{+}
ShovFP	85.0	85.5	80.0^{-}	85.5

complexity. Complexity of classification datasets is extensively studied in [8] and [5] where several measures of dataset complexity were introduced. The authors divide their complexity measures into 3 categories: measures of overlap of individual feature values, measures of separability of classes, measures of geometry, topology and internal density of manifolds. We adopt complexity measures from the first two categories which we explain in the following sections.

5.1 Average Terms per Document

We consider the number of terms per document an important characteristic of datasets under the hypothesis that the average length of documents is a good indicator of the performance of term relatedness. The number of terms in a document is calculated after text preprocessing: stopwords removal, term normalisation and feature selection. Thus the count of terms in document is restricted to the terms from the indexing vocabulary. The average term count for the entire dataset is calculated by taking the average term count for all documents in the dataset as in equation 6.

$$AveTermCount = \frac{\sum_{d_i \in D} TermCount(d_i)}{|D|} \tag{6}$$

5.2 Measures of Overlap of Individual Feature Values

Metrics in this category estimate the complexity of datasets by the existence of one or more highly discriminatory features. The one metric we use from this category is **Fisher's Ratio** which is given in equation 7.

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{7}$$

where μ_1, μ_2, σ_1^2, σ_2^2 are the means and variances of the two classes respectively. Fisher's ratio is defined for a single feature while text representations are typically multidimensional. However, as long as there exists one highly discriminatory feature, the dataset can be considered to have low complexity. Therefore, the maximum ratio over all the feature dimensions is selected. High values of this measure indicate that the dataset has low complexity due to the presence of highly discriminatory features and vice versa.

5.3 Measures of Separability of Classes

Measures in this category estimate the distance between documents in the same class in comparison with the distance between documents in different classes. Standard distance/similarity metrics can be used for this purposes. Here, we use cosine similarity to measure the distance between documents.

Average Intra/Inter class NN Distance Ratio This measure calculates the average ratio of the distance of each document from its nearest neighbour of the same class (intra-class distance) and its nearest neighbour of a different class (inter-class distance). This is show in equation 8.

$$N_2 = \frac{\sum_{d \in D} IntraClassDist(d)}{\sum_{d \in D} InterClassDist(d)} \tag{8}$$

Low values of this measure suggest that the distance between documents of the same class is smaller than their distances to documents of a different class which in turn indicates low complexity.

1NN Error Rate This measure is computed by calculating the leave-one-out error rate of a one-nearest-neighbour classifier. Again we use cosine similarity here to determine the nearest-neighbour. A high error rate suggests that documents are very similar across class boundaries which indicates high complexity.

Complexity Profile This measure was originally proposed for measuring the complexity of a case base by looking at the classes of the nearest neighbours of each case [10]. This is calculated by iteratively retrieving successively larger neighbourhoods k of a document d up to the neighbourhood size K (we use $K = 10$ in this work). This is shown in equation 9.

$$Complexity(d) = 1 - \frac{\sum_{k=1}^{K} P_k(d)}{K} \tag{9}$$

$$P_k(d) = \frac{\sum_{i=1}^{k} \phi(n_i)}{k} \tag{10}$$

Where n_i is a document in the neighbourhood of d and

$$\phi(n_i) = \begin{cases} 1, & \text{if } d \text{ and } n_i \text{ belong to the same class} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

For each neighbourhood size k, equation 9 computes the proportion $P_k(d)$ of documents in that neighbourhood that belong to the same class as d. The values of $P_k(d)$ for $k = 1...K$ are then averaged and the average is subtracted from 1. The final complexity measure for the entire dataset is computed as the average complexity of all documents as in equation 12

$$CompProfile = \frac{\sum_{d \in D} Complexity(d)}{|D|} \tag{12}$$

Weighted Complexity Profile Complexity profile fails to take into account the distances of neighbours from d when computing $P_k(d)$. This means that all neighbours are considered equal even though some may be much closer to d than others. Thus, the similarity-weighted complexity profile addresses this limitation by assigning weights to neighbours according to their similarity to d as in equation 13.

$$P_k(d) = \frac{\sum_{i=1}^{k} \phi(n_i) sim_{d,n_i}}{k} \tag{13}$$

From equation 13, neighbours that are closer to d contribute more to the computation of $P_k(d)$ than more distant neighbours. Thus a higher value of $P_k(d)$ not only indicates that many neighbours of d belong to the same class as d, but also these neighbours are within close proximity to d. The complexity profile for the entire dataset is computed as in equation 12.

5.4 Measures of Attribute Distribution

Because of the high dimensionality of text representations, the distribution of attributes between classes is a more important measure of dataset complexity than the overlap of individual attribute values. Thus in this section, we introduce new measures of classification dataset complexity. Our measures are based upon feature selection metrics. Feature selection metrics assign a score to each term in the vocabulary which represents the value or informativeness of the term. Intuitively, we can consider a dataset with many high value terms to have low complexity as it would be easier to classify, and vice versa. Accordingly, we

present six new complexity measures using standard feature selection algorithms.

Chi^2 The Chi^2 metric measures the lack of independence between two variables. Given a two way contingency table of a term t and a class c, let N be the total number of documents in the dataset and $f(t, c)$ be a function that returns the co-occurrence count of a term t and a class c. The Chi^2 score of t can be measured as:

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \tag{14}$$

where:

$A = f(t, c)$, $B = f(t, \bar{c})$, $C = f(\bar{t}, c)$ and $D = f(\bar{t}, \bar{c})$.

Accordingly, we crate two complexity measures: **Average Chi Score** which is the mean chi^2 score over all terms and **Maximum Chi Score** which takes the maximum chi^2 score over all terms.

Information Gain Information Gain (IG) measures the information obtained for category prediction by knowing the presence or absence of a term in a document. Let n be the number of categories, the IG of a term t can be calculated as:

$$IG(t) = -\sum_{i=1}^{n} P(c_i) log P(c_i) + P(t) \sum_{i=i}^{n} P(c_i, t) log P(c_i, t) + P(\bar{t}) \sum_{i=i}^{n} P(c_i, \bar{t}) log P(c_i, \bar{t}) \tag{15}$$

Using the IG metric, we create two complexity measures: **Average IG Score** which is the mean IG score over all terms and **Cumulative IG Score** which is the sum of IG scores for all attributes.

Mutual Information The Mutual Information (MI) metric compares the probability of observing two variables together with the probability of observing them independently. Thus given a term t and a class c, the mutual information of between them is measures by:

$$MI(t, c) = log_2 \frac{P(t, c)}{P(t)P(c)} \tag{16}$$

Similarly, we create two complexity measures using the MI metric: **Average MI Score** and **Cumulative MI Score**.

6 Predicting When to Generalise

The aim of this evaluation is to determine how well we can predict when to and when not to use generalisation for text classification. We use a case based approach where we create a separate casebase for each term relatedness technique. Note that our aim is not to compare between the different approaches for extracting term relatedness. Rather, given any useful measure of term relatedness, can we determine when to apply such to a given dataset.

Each case in a casebase represents a single dataset. The description of a case corresponds to the set of features described in Section 5 and the case solution is a binary judgement of whether or not to use generalisation for that dataset. A case is labelled with the decision to use generalisation if the particular term relatedness technique being considered has improved text classification accuracy by at least 1%. Otherwise, we label the case with the decision not to use generalisation. For example for the Cooc technique, generalisation produced an improvement of 1.5% compared to Base on the Hardware dataset (see table 1) and the decision to use generalisation is selected as the case solution for Hardware. On the other hand for the MedSpace dataset, Cooc produced a decline of 2.3 % in classication accuracy and thus the solution for this case in not to use generalisation.

The similarity between cases is determined using the average interval similarity defined for real values attributes in the JColibri framework [12]. The average interval similarity of two cases a and b with N attributes is given in equation 17.

$$Sim(a,b) = \frac{\sum_{i=1}^{N}(1 - |a_i - b_i|)}{N} \tag{17}$$

We report the classification accuracy over a leave-one-out validation using a 1-NN approach. Our baseline is a zeroR approach that always selects the majority class i.e. apply generalisation.

Table 2. Classification accuracy of generalisation prediction.

	Cooc	Npmi	LSI
ZeroR	58.0	51.0	52.0
kNN	74.4	72.1	81.4

From the results shown in table 2 , it is clear that our case-based approach out performs zeroR. This means that our set of features are predictive of the effectiveness of applying generalisation for text classification. Our analysis of the dataset attributes indicates that improvement of classification performance after generalisation is correlated with dataset complexity. For example the Pearson's correlation scores between the NN Error Rate measure and the difference in classification accuracy with baseline for Cooc, Npmi and Lsi respectively are 0.52, 0.32 and 0.54 which indicate strong positive correlation. This means that

the more complex a dataset is, the more useful term relatedness is for improving classification performance.

7 Related Work

Much work has been done in the area of meta-learning. For example the approach presented in [1] uses a meta learner to assign classifiers to datasets. Of particular relevance to our work are meta case-based approaches e.g. [6] where a meta case-based technique is used for selecting case-base maintenance algorithms. In this approach, an individual meta-case models a entire case-base where the case solution is the maintenance algorithm that provides the best performance on that case-base and the case description comprises a set of attributes that are derived using complexity measures. The system was evaluated on a set of 25 classification datasets with promising results. Although our approach uses many of the same features as [6], our tasks are different. Our approach is concerned with predicting when to generalise while the approach in [6] is concerned with selecting the best algorithm to use for case-base maintenance.

Another approach that uses CBR for selecting the best sentiment lexicon given a sentiment classification dataset is presented in [11]. Here also, a dataset is represented as a single case where the case solution is the best performing sentiment lexicon for the dataset. The case description is modelled as an n-dimensional feature vector derived from document, sentence and term-level statistics of as well counts of part-of-speech information and punctuations. The features chosen for case representation are designed to capture the subjectivity of the corresponding dataset. On the other hand, our approach is concerned with trying to predict, given any generic text classification dataset, when to apply generalisation for text representation. Consequently, we use a totally different set of features from [11].

The system presented in [9] uses a CBR approach to select the best classification algorithm for a dataset. The datasets considered here are not limited to textual datasets and the features used for case representation are designed to capture characteristics of datasets that contain both numeric and symbolic attributes. As such, these features are different from the ones used in our approach and are perhaps not optimal for textual datasets.

8 Conclusion

In this work we have discovered that generalisation is not always beneficial for text classification by investigating the performance of 3 different co-occurrence based term relatedness techniques on 43 text classification datasets. Accordingly we presented a case-based approach for predicting when to generalise. Our case description was obtained from several statistical metrics that capture different attributes of classification datasets. Results show that our case-based approach outperforms a zeroR baseline over all three term relatedness techniques. These results indicate that our case-based approach is able to correctly predict the

performance of generalisation on a range of text classification datasets. Furthermore, our analysis of the dataset attributes indicates that the performance of generalisation is correlated with dataset complexity.

References

1. Bensusan, H., Giraud-Carrier, C., Kennedy, C.: A higher-order approach to meta-learning. In: Proceedings of the ECML'2000 workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination. pp. 109–117. ECML'2000 (June 2000)
2. Brants, T., Inc, G.: Natural language processing in information retrieval. In: In Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands. pp. 1–13 (2004)
3. Chakraborti, S., Wiratunga, N., Lothian, R., Watt, S.: Acquiring word similarities with higher order association mining. In: Weber, R., Richter, M. (eds.) ICCBR. LNCS (LNAI), vol. 4626, pp. 61–76. Springer (2007)
4. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. 16(1), 22–29 (Mar 1990)
5. Cummins, L., Bridge, D.: On dataset complexity for case base maintenance. In: Ram, A., Wiratunga, N. (eds.) Case-Based Reasoning Research and Development (Procs. of the 19th International Conference on Case-Based Reasoning). pp. 47–61. LNAI 6880, Springer (2011)
6. Cummins, L., Bridge, D.: On dataset complexity for case base maintenance. In: Ram, A., Wiratunga, N. (eds.) Case-Based Reasoning Research and Development (Procs. of the 19th International Conference on Case-Based Reasoning). pp. 47–61. LNAI 6880, Springer (2011)
7. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
8. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 289–300 (2002)
9. Lindner, G., Ag, D., Studer, R.: Ast: Support for algorithm selection with a cbr approach. In: Recent Advances in Meta-Learning and Future Work. pp. 418–423 (1999)
10. Massie, S., Craw, S., Wiratunga, N.: Complexity profiling for informed case-base editing. In: Proceedings of the 8th European conference on Advances in Case-Based Reasoning. pp. 325–339. ECCBR'06, Springer-Verlag, Berlin, Heidelberg (2006)
11. Ohana, B., Delany, S., Tierney, B.: A case-based approach to cross domain sentiment classification. In: Agudo, B., Watson, I. (eds.) Case-Based Reasoning Research and Development, Lecture Notes in Computer Science, vol. 7466, pp. 284–296. Springer Berlin / Heidelberg (2012)
12. Recio-García, J.A., Sánchez-Ruiz, A.A., Díaz-Agudo, B., González-Calero, P.A.: jcolibri 1.0 in a nutshell. a software tool for designing cbr systems. In: Petridis, M. (ed.) Proccedings of the 10th UK Workshop on Case Based Reasoning. pp. 20–28. CMS Press,University of Greenwich (2005)
13. Sani, S., Wiratunga, N., Massie, S., Lothian, R.: Term similarity and weighting framework for text representation. In: Proc. of the 19th International Conference on Case-Based Reasoning, ICCBR (2011)