Enhancing Short Text Clustering with Small External Repositories

Henry Petersen **Josiah Poon**

School of Information Technologies,
University of Sydney, NSW 2006, Australia
Email: {hpet9515,josiah}@it.usyd.edu.au

Abstract

The automatic clustering of textual data according to their semantic concepts is a challenging, yet important task. Choosing an appropriate method to apply when clustering text depends on the nature of the documents being analysed. For example, traditional clustering algorithms can struggle to correctly model collections of very short text due to their extremely sparse nature. In recent times, much attention has been directed to finding methods for adequately clustering short text. Many popular approaches employ large, external document repositories, such as Wikipedia or the Open Directory Project, to incorporate additional world knowledge into the clustering process. However the sheer size of many of these external collections can make these techniques difficult or time consuming to apply.

This paper also employs external document collections to aid short text clustering performance. The external collections are referred to in this paper as Background Knowledge. In contrast to most previous literature a separate collection of Background Knowledge is obtained for each short text dataset. However, this Background Knowledge contains several orders of magnitude fewer documents than commonly used repositories like Wikipedia. A simple approach is described where the Background Knowledge is used to re-express short text in terms of a much richer feature space. A discussion of how best to cluster documents in this feature space is presented. A solution is proposed, and an experimental evaluation is performed that demonstrates significant improvement over clustering based on standard metrics with several publicly available datasets represented in the richer feature space.

keywords: Text Mining, Clustering, Short Text, Background Knowledge

1 Introduction

The huge volume of information available through resources such as the world wide web has driven much interest in the clustering and automated analysis of textual data. Most algorithms represent text using a model derived from a bag-of-words. In the bag-of-words model a single feature is created for each word in the corpus and each document is assigned an attribute value for that feature corresponding to the number of occurrences of that word the document.

A fundamental requirement for effective text clustering algorithms is the ability to compare documents according to their semantic content. However for such tasks the use of a bag-of-words representation introduces a number of problems. For realistic document collections, vocabulary sizes in the tens or even hundreds of thousands are not uncommon which can lead to a feature space that is highly sparse. On top of this, issues such as synonymy (different words used to denote the same concept) and polysemy (a single word that can denote multiple concepts) can further degrade the ability of an algorithm to successfully analyse a text collection. These issues are even more pronounced when the text being analysed comprises short strings (ie. documents containing perhaps only a few words each).

A number of researchers have made use of external knowledge repositories in an attempt to extract additional information and compensate for the sparsity of the feature space. Previous literature has reported a wide range of sources for gaining such external knowledge including search engines such as Google (Sahami & Heilman (2006)) as well as linguistic resources like Wordnet (Hotho et al. (2003)).

Recent work in both the supervised and unsupervised literature has explored obtaining external knowledge from large static repositories of text like Wikipedia (Gabrilovich & Markovitch (2007), Hu et al. (2009)) and the Open Directory Project (Gupta & Ratinov (2008)). These approaches have achieved significant success, however they are not without drawback. Because of the very large size of the collections these techniques can be quite time consuming to apply (Phan et al. (2008)). Additionally it has been claimed that for a given text analysis task some consistency between the topic structure of the external knowledge repository and the text collection being analysed (referred to from now on as the target collection) is required in order for the external knowledge to be effective (Phan et al. (2008)). While repositories such as Wikipedia have been shown to be effective for many tasks they are unlikely to be as effective for highly technical or specific problem domains. For such domains the collection of an appropriate, suitably large corpus for use as an external knowledge repository may not always be straightforward.

Some work within the supervised classification literature (Zelikovitz & Hirsh (2001), Zelikovitz & Hirsh (2002), Weng & Poon (2006)) has explored using much smaller external repositories of unlabelled text. These smaller external collections are referred to as Background Knowledge. Their work uses the Background Knowledge to map short text strings into an alternative representation called the Bridging space.

The Background Knowledge need not be drawn from the same distribution as the short text documents, and can differ significantly in length and structure. Additionally, and in contrast to much of the supervised literature the size of the Background Knowledge collection is very small, often only one or two thousand documents in total. However the methods proposed by Zelikovitz et al. for use of this Background Knowledge are specific to supervised classification, and to the best of our knowledge no previous application of such Background Knowledge to document clustering exists.

The following is a brief summary of several interesting contributions provided in this paper:

- Previous clustering literature has often used very large collections such as Wikipedia and the ODP. We demonstrate that small external document collections can still substantially increase short text clustering performance. Furthermore the methods used to exploit these small collections need not be particularly complex.

- To the best of our knowledge all previous applications of Background Knowledge and the Bridging space have focused on supervised classification problems (Weng & Poon (2006), Zelikovitz & Hirsh (2002)). We demonstrate the effectiveness of the Bridging space for short text clustering tasks.

- A function is proposed for use clustering text represented in a Bridging space created using the Background Knowledge. The cluster purity obtained using the proposed function in this feature space is demonstrated experimentally to be substantially better that obtained using several standard similarity functions including the cosine and euclidean distance.

2 Background Knowledge

We now present an explicit definition of Background Knowledge as it is used in this work. An item of Background Knowledge can be any text document semantically relevant to the problem domain of the target corpus. These text documents are unlabelled and no requirement is made that the target and Background documents be drawn from the same distribution. Background documents may be substantially greater in length than the short text in the target collection. The only requirement is that the Background documents be semantically related to the target domain.

This requirement means that identifying an appropriate source of Background Knowledge is a problem specific task. For example given a target collection of short text consisting of a set of technical paper titles we wished to compare according to sub-discipline, the text documents used to create the Background Knowledge could be abstracts or full text from similar papers, excerpts from text books, or even text from relevant mailing lists or forums.

The Background Knowledge collections used in this work are also in general quite small (at most several thousand documents - see section 5.1). This is in contrast to many other algorithms involving static text repositories where the number of external documents is frequently in the hundreds of thousands or even millions (Gabrilovich & Markovitch (2006), Phan et al. (2008)).

2.1 Motivation

Due to the highly sparse nature of short text, it can be very difficult for algorithms to effectively model the co-occurrence structure of a short text collection. Within such a problem domain, it is highly likely that many related words will exist that might never occur together. For example the words 'computer', 'laptop', 'pc' and 'notebook' are all semantically related, however when dealing with short text corpora are unlikely to appear together in a single document.

Each individual Background document in the corpus of Background Knowledge is drawn from one or more latent topics relevant to the clustering task at hand, and the words they contain will be drawn from these topics. Therefore, given a collection of larger Background documents suitably drawn from latent topics semantically related to the problem domain, there is a good possibility that such a pair of related words from the short text will co-occur in the larger documents (Zelikovitz (2002)). As a results of this, additional information on the co-occurrence structure of the problem domain can be obtained from the Background Knowledge (Zelikovitz (2002)).

2.2 Bridging Space

In order to utilise the additional co-occurrence information in the Background Knowledge we map the target documents into a much richer space that will facilitate better comparison between each target instance. This alternate representation is referred to as the 'Bridging Space' (Weng & Poon (2006)). In order to represent a short text in the Bridging space we generate one feature for each document in the Background corpus, and assign to each feature an attribute value equal to the result of the cosine similarity between the short text and the corresponding Background document. More explicitly, given a vector x describing a target document, and a collection of N items of Background Knowledge $B = \{b_1, b_2, \ldots, b_N\}$ expressed over an identical vocabulary, the target document represented in the Bridging space is defined as follows[1]:

$$\hat{x} = \{\hat{x_1}, \hat{x_2}, \ldots, \hat{x_N}\}$$

$$\hat{x}_i = \frac{x \cdot b_i}{\|x\| \times \|b_i\|}$$

Despite two short strings potentially having no common terms, any semantic relationships between them are far more likely to be apparent in the Bridging space. This is because related terms from the two strings are likely to occur together somewhere in the Background Knowledge, producing similar values for the features corresponding to the Background documents in which they co-occur. It is unlikely however that each Background document will describe only a single latent topic. As Background documents almost certainly contain terms drawn from separate (although likely still related) topics, the potential for links based on terms drawn from separate topics may introduce noise when comparing text using the Bridging space. In practice, this effect appears to be is mitigated by selecting a corpus of Background Knowledge sufficiently relevant to the target domain, and by ensuring it contains enough Background documents to

[1] In this work, comparisons between x and each b_i are made using the cosine similarity function. Although we do not do so, other functions could of course be used.

$$euclidean = \sqrt{\sum |x_i - y_i|^2}$$

$$cosine = \frac{x^T \cdot y}{\|x\| \times \|y\|}$$

$$extendedjaccard = \frac{x^T \cdot y}{\|x\|^2 + \|y\|^2 - x^T \cdot y}$$

Figure 1: Several common similarity and distance functions for two vectors x and y

overcome the noise (the ideal number of Background documents is likely to be specific to the domain at hand, however further investigation is left for future work).

The Bridging space has been used several times previously within the literature on supervised learning (Zelikovitz & Hirsh (2001), Zelikovitz & Hirsh (2002), Zelikovitz & Hirsh (2005), Chan et al. (2006), Weng & Poon (2006)). The work presented in this paper is novel however in that it describes the first application of the Bridging space to unsupervised tasks (several of the prior approaches could be adapted to unsupervised learning but their use in such scenarios is potentially sub-optimal. These methods are discussed in more detail in section 4). Additionally to the best of our knowledge there has been no previous treatment of how documents represented in the Bridging space should be compared. We provide such an analysis in the following section.

Finally although it is the case for all Background collections used in this work there is no reason to indicate that Background Knowledge needs to be formed from documents drawn from a single source. In fact such Background Knowledge has previously been used effectively in the supervised literature concerning the Bridging space (Zelikovitz & Kogan (2006)) although further examination of such Background Knowledge for our purposes is left for future work.

3 Clustering in the Bridging Space

When clustering a collection of text, we typically try to find a solution that maximises the intra-cluster and minimises inter-cluster values of some measure of relatedness between instances over the entire data. Within the literature, a wide range of similarity and distance functions have been employed for this purpose with a variety of applications. Several widely used functions are shown in Figure 1. It is well known that certain functions produce better performance with different types of data (Joydeep et al. (2000)). This section presents a discussion of what properties a good function for measuring how well two instances belong in the same cluster should possess. We then propose a clustering function that, when applied over a clustering of short texts represented in the Bridging space will provide a good measure of the overall quality of the solution.

Given a short text collection and associated corpus of Background Knowledge, each Background document can be said to describe some combination of latent concepts from the problem domain. Each document may describe multiple concepts and an individual latent concept may also appear in multiple Background documents. Representing a text document in the Bridging space then describes that snippet in terms of a set of similarities between the snippet and groups of latent concepts.

While the bridging space used in this work is similar in many ways to previous approaches such as Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch (2006), Gabrilovich & Markovitch (2007)) there are some differences due to the type of external text corpora employed. For example, in their paper on ESA, Gabrilovich & Markovitch (2007) use Wikipedia articles to form the Background Knowledge. They note that each article will describe a single topic, each of which is "Explicitly defined and described by humans", which is not the case for the work described in this paper. Each Background document we employ may reference several latent concepts, which in turn may appear in one or more documents throughout the Background Knowledge corpus. The requirements for an optimal clustering function when using Background Knowledge may indeed be different therefore to such previous work.

We start by stating the following desirable property that a good clustering function for use in the Bridging space should possess:

Proposition 1: *The value produced by the function for any pair of short texts should depend only on features with non-zero similarity to both documents (ie. the short texts both share at least one term with the corresponding Background document).*

As previously stated, each Background document is understood to describe a combination of latent semantic topics from the problem domain. It is reasonable to assume that for a given corpus of Background Knowledge some topics will occur more frequently than others. If a given set of topics is over represented in the Background Knowledge then a comparison function that considers features with zero similarity may adversely effect comparisons between snippets unrelated to those topics. In this way using such a comparison function helps compensate for topic bias in the Background Knowledge.

Additionally, given that a single Background document is unlikely to contain all terms relevant to the topics it describes, a similarity function conforming to proposition 1 has the added benefit of guarding against a short string being erroneously adjudged as not related to a relevant topic.

None of the standard comparison functions described in Figure 1 conform to proposition 1. For example when comparing two snippets the Euclidean function (which is based on the difference between attribute values) will treat Background documents sharing no terms with the two snippets identically to Background documents that share many with both. This clearly violates the desired property that Background documents sharing no common terms with either snippet be discarded. The cosine similarity and extended jaccard coefficient also do not have this property as can be seen from their equations in Figure 1; a Background document that shares terms in common with only one snippet will increase the length of the vectors (and therefore the denominators) with out increasing the dot-product (leaving the numerator unchanged).

We now present a proposed function for use within the Bridging space that contains the aforementioned desirable property. Let x and y denote two vectors describing strings represented in a Bridging space with N features, and x_i and y_i reference the i^{th} attribute of the strings x and y respectively. We then compute the vector corresponding to the the element wise product of x and y:

$$EWP = \{ewp_1, ewp_2, \ldots, ewp_N\}, ewp_i = 1 - x_i \cdot y_i \tag{1}$$

Let π define an ordering on EWP such that:

$$ewp_{\pi(i)} \leq ewp_{\pi(j)} \forall i < j$$

For two documents x and y, the proposed clustering function is then defined as:

$$bridge(x, y) = 1 - \prod_{i=1}^{k} ewp_{\pi(i)} \tag{2}$$

In order to satisfy proposition 1 the proposed function operates on the product of individual attribute values. Observe that the vector computed in (1) will produce a value of 1 iff either x or y are 0 and a value in the range [0,1) otherwise (assuming of course that both x and y are in the range [0,1]). Note also that the bridge function presented in (2) does not employ an operator such as the sum, but instead uses the product operator to combine the values. Because of this, proposition 1 is upheld as the values of one in EWP will not impact the computed score.

Note that the sum operator is not appropriate in this case. For two document vectors, the sum of the product of the attribute values is proportional to the angle between the vectors. However in order to satisfy property one we consider only the features for which both x and y are non-zero. Because of this, in practice the angle between the two vectors would be very small and lead to much less variation in values between different pairs of documents.

The proposed function also contains several other desirable properties. The value k in equation (2) controls the number of Background documents that are taken into account in measuring similarity between pieces of text. Recall that some topics can be expected to occur more frequently in the Background Knowledge than others. If this imbalance is large enough, it is possible that the results of comparisons related to these topics could be unfairly inflated. Capping the number of Background documents in this manner helps to introduce some tolerance for this problem; only the top k links through the Background Knowledge are considered which helps to alleviate over-representation of topics in the corpus. In practice, the value used for k is identified as a parameter of the algorithm.

The proposed function will also produce lower values for pairs of snippets that share terms with less than k Background documents. Consider two snippets x and y that have a cosine similarity of 0.2 to 10 Background documents. The function presented in 2 will return a similarity value for these snippets of 0.8926. However if x and y have a cosine similarity of 0.2 to only 5 Background documents, the function will return 0.6723. This property is desirable as having a larger number of links through the Background Knowledge provides confidence that two snippets are in fact related and the result is not merely anomalous.

Choosing a good value for the parameter k involves ensuring a reasonable number of Background documents will be considered while not unfairly penalising snippets related to rarer but still relevant topics in the Background Knowledge. In this work we use a k value 10 for all experiments. We note however that the optimal value will likely depend on the target collection and Background Knowledge, although a more detailed evaluation is left for later work.

$$M = \left(\begin{array}{ccc} d_1 & d_2 & d_3 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right)$$

$$BG = \left(\begin{array}{ccc} b_1 & b_2 & b_3 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array} \right)$$

$$\hat{M} = \left(\begin{array}{ccc} d_1 & d_2 & d_3 \\ \frac{\sqrt{3}}{2} & 0 & \frac{3}{4} \\ 0 & \frac{\sqrt{3}}{2} & \frac{1}{2} \\ \frac{1}{3\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{1}{2\sqrt{6}} \end{array} \right)$$

Figure 2: Sample term document matrices for target and Background document collections, along with the target collection represented in the Bridging space

By maximising the total combined values of (2) between elements in each cluster, we expect to generate clusters with a high likelihood of describing similar sets of latent topics from the Background Knowledge. However, when considered purely as a direct measure of similarity between pairs of short texts, (2) appears to possess some interesting properties. When used as a measure of distance, (2) is non-metric as it does not obey the triangle inequality, nor is the distance between an element and itself guaranteed to be 0. To prove these claims, consider figure 2 which shows the attribute values for three short text snippets as well as a Background Knowledge corpus with three documents. Figure 3 shows the EWP vectors for each pair of documents along with a distance matrix corresponding to 1 minus the values of 2 for each pair of documents using a k value of 1.

Theorem 1. *When treated as a measure of distance, the proposed function does not obey the triangle inequality.*

Proof. The value produced by the proposed function between documents d_1 and d_2 is 0.9444. This value is greater than the sum of values between documents d_1 and d_3 and documents d_2 and d_3 $(0.3505 + 0.5670 =$

$$EWP = \left(\begin{array}{cccc} & b_1 & b_2 & b_3 \\ d_1, d_1 & 0.25 & 1 & 0.9444 \\ d_1, d_2 & 1 & 1 & 0.9444 \\ d_1, d_3 & 0.3505 & 1 & 0.9519 \\ d_2, d_2 & 1 & 0.25 & 0.9444 \\ d_2, d_3 & 1 & 0.5670 & 0.9519 \\ d_3, d_3 & 0.375 & 0.75 & 0.9583 \end{array}\right)$$

$$OneMinusBridge = \left(\begin{array}{cccc} & d_1 & d_2 & d_3 \\ d_1 & 0.25 & 0.9444 & 0.3505 \\ d_2 & 0.9444 & 0.25 & 0.5670 \\ d_3 & 0.3505 & 0.5670 & 0.375 \end{array}\right)$$

Figure 3: EWP vectors and similarity matrix for Figure 2

0.9175 < 0.9444), and thus the proposed distance function does not obey the triangle inequality. □

The notion of semantic similarity has long been known to not follow the triangle inequality (Tversky (1977)). We can observe this by considering the similarity between the terms *Trees*, *Flowers*, and *Chocolates*. It can be seen that the word pairs (*Trees*, *Flowers*) and (*Flowers*, *Chocolates*) have a high similarity (they describe plants and gifts respectively), however this is not the case for the pair (*Trees*, *Chocolates*). That the proposed function does not obey the triangle inequality is therefore not a problem.

Theorem 2. *The distance between an element and itself is not always 0. Furthermore it may not be minimal.*

Proof. From the matrix in 3, the values between documents d_1, d_2 and d_3 and themselves are respectively 0.25, 0.25 and 0.375. It can therefore be seen that when using the proposed function as a measure of distance, the distance from an element to itself will not necessarily be 0. That the distance between an element and itself can be greater than it is to another element can be demonstrated by observing that $bridge(d_1, d_3) = 0.3505 < 0.375 = bridge(d_3, d_3)$. □

Recall that each Background document describes a number of latent topics from the problem domain, and that when represented in the bridging space an individual attribute can be considered to describe the similarity between a short text string and the topics described in a piece of background knowledge. It follows then that when applied to two short text snippets, the function described in (2) can be regarded as a combination of the similarities between the two snippets and the set of latent topics described in k pieces of background knowledge. In other words, rather than producing a value that directly compares the two snippets, the function will produce a value comparing their similarity to some set of latent topics.

Consider the example from Figure 2 discussed above in Theorem 2. All three terms in d_1 are present in the background document b_1, which implies that it is very likely d_1 is related to the latent topics described in b_1. The short string d_3 also shares three terms with b_1, however d_3 also contains an additional term not contained in b_1. This could be considered to reduce the likelihood that d_3 is related to the latent topics described in b_1. The additional term means that the similarity between d_3 and b_1 will be less that that of d_1 and b_1. This leads to the function in (2) indicating d_1 and d_3 have a greater similarity to a common set of topics in the Background Knowledge than it does with d_3 and d_3.

Although these properties imply that the clustering function proposed in (2) is less than ideal for directly comparing individual pairs of short strings, it is still reasonable to produce a clustering based on optimising (2) over a collection of documents. Although local discrepancies may exist, such a clustering would still tend to group documents related to similar latent topics. Providing the Background Knowledge adequately describes an appropriate set of latent topics, the effect of any local inconsistencies should be outweighed. In the following sections we demonstrate the effectiveness of clustering using (2) in the Bridging space compared to optimising clusters based on other functions.

The proposed function in 2 bears some similarity to the supervised Bridging algorithm (Zelikovitz & Hirsh (2002)) (see section 4 for further detail). Again however, the supervised Bridging algorithm relies on the presence of labelled training data to function, and as opposed to comparing text directly it instead measures similarity between text and each class label in the data set.

4 Related Work

Clustering short text based on semantic similarity is a problem that has seen much interest in recent times. A wide range of approaches have been proposed to compensate for the difficult nature of the task which in turn can be broadly divided into two categories; internal and external.

Internal analysis techniques are those that attempt to discover the semantic relationships between individual terms through statistical analysis of the target document collection. They consider no additional knowledge repositories. This class of approach includes techniques like Latent Semantic Indexing (LSI)(Deerwester et al. (1990)), Probabilistic Latent Semantic Analysis (pLSA) (Hofmann (1999)), and Latent Dirichlet Allocation (Blei et al. (2003)). While these approaches have proved successful in many cases, their effective application can be difficult where the target collection is extremely sparse or there are insufficient instances to adequately model the problem domain. For brevity's sake we do not discuss them further here and the interested reader is directed to the appropriate literature.

The second class of solution for measuring semantic similarity between pairs of short text involves the application of additional data not available in the original dataset (hence we refer to such approaches as external).

A popular approach within the literature has been the application of lexical resources such as WordNet (Miller (1995)) to aid in the comparison of textual data. Wordnet provides a manually annotated lexical database of the English language, and was originally created by G. Miller at Princeton university. By taking advantage of the semantic relationships expressed between terms in Wordnet, several methods have been proposed for compensating issues of semantic ambiguity when comparing text (Hotho et al. (2003), Jing et al. (2006), Li et al. (2008)). One drawback to these methods is that the creation and maintenance of such resources can be very expensive, and obtaining a suitable resource may be difficult for some domains.

Several previous works (ie. Sahami & Heilman

Table 1: Summary of the datasets used.

Dataset	# Docs	Avg. Doc. Length	# Classes	# BG Docs.	Avg. BG. Doc. Length
2CNews	1033	6.02	2	1165	56.92
2CPhys	953	5.82	2	1531	74.46
3CPhys	1066	5.80	3	1702	72.69
7CNetv	1723	2.53	7	1160	103.31

(2006), Yih & Meek (2007), and Bollegala et al. (2007)) propose methods to employ the results of Google searches on short text strings to measure their similarity. While such algorithms have proven effective for suitably short text, they are inappropriate for application to longer documents. This is due to the algorithms' use of the target short text snippet as Google queries. Our approach has no such limitation.

Some researchers (ie. Gabrilovich & Markovitch (2007), Banerjee et al. (2007), Hu et al. (2008), Hu et al. (2009), and Phan et al. (2008)) have made use of large static repositories of text such as Wikipedia and the Open Directory Project (ODP). One issue with such methods however is that the sheer number of documents in these collections (often hundreds of thousands or even millions) can lead to serious issues with regards to the processing time required (Phan et al. (2008)). As well as this many of the more general sources like Wikipedia are unlikely to be optimal for highly technical domains, and finding a suitably large corpus for such problems is not straightforward. This class of approach shares some similarities with the one presented in this paper as both employ static text collections to obtain additional domain knowledge, although ours differs in that that the collections used are of a much smaller size.

Within the clustering literature work exists where additional domain knowledge is added through placing constraints on the output of the algorithm. For example, in the paper by Wagstaff et al. (2001) an expert user is employed to annotate part of the target dataset with must-link and cannot-link constraints, and an extension to the k-means algorithm is provided to utilise this information and significantly improve the resultant clustering. These approaches differ from the other external methods presented above in that they do not address the problem of measuring similarity between individual instances, rather they modify the algorithm applied to analyse a given dataset. Such methods suffer from the fact that the manual annotation by an expert user required for their application is not always feasible, even if such an expert user exists.

The Bridging space used in this work has been seen previously in the supervised text classification literature. The Bridging space was first used implicitly in the work of Zelikovitz & Hirsh (2001) in their nearest neighbour Bridging approach. Given a set $(t_i, l_i) \in Tr$ of N labelled training documents, a set $b_j \in Bg$ of M Background documents, a query string q, a similarity function sim and an integer value k the nearest neighbour Bridging algorithm works by first computing the length $N \times M$ vector of 2-tuples:

$$\{(1 - sim(t_i, q) \times sim(b_i, q), l_i)\}$$

The tuples are then ordered increasing according to their first value, and all but the first k are discarded. Each class is then assigned a score equal to one minus the product of the first value for all remaining tuples with that class label. The query is finally assigned the label corresponding to the largest such value[2].

The supervised Bridging algorithm (Zelikovitz & Hirsh (2002)) is in some respects very similar to the approach proposed in this work. The Bridging algorithm however is only suitable for supervised tasks, as it at no time computes a single similarity value between individual text strings (rather it computes similarities between instances and classes).

The Bridging space has also been used in the literature on classifying imbalanced data (Weng & Poon (2006)). Documents are expressed in the Bridging space defined by the Background Knowledge before standard classification algorithms (in their case a support vector machine) are applied. The work described in by Weng et al. differs from ours in that we propose a method for comparing documents in the Bridging space as an alternative to standard algorithms.

The Bridging space has also been used previously (although it is not referred to as such) in Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch (2007)) using Wikipedia articles as Background Knowledge. Our work differs in that, like Weng & Poon (2006), the approach in Gabrilovich & Markovitch (2007) compares snippets in the Bridging space using standard similarity metrics. Additionally ESA employs a very large, general repository of Background Knowledge (Wikipedia), while our work employs much smaller, targeted document collections.

Further interesting work using the Bridging space has been performed in Chan et al. (2006) by modifying the Bridging algorithm of Zelikovitz et al. for use with Background Knowledge of the same size and form as the target text snippets (eg. Semi-supervised learning). In their work the authors note a slight deterioration in the performance of the Bridging algorithm when used with Background Knowledge of this type which can be attributed to the decreased generality of the Background documents and the associated domain knowledge they provide. The authors show however that this effect can be offset by employing the Bridging algorithm in conjunction with semi-supervised techniques such as co-training (Blum & Mitchell (1998)) and assigning labels to a portion of the Background Knowledge.

Finally we note one prior use of this type of Background Knowledge that could be applicable to clustering tasks, however our method improves upon this in several ways. Zelikovitz & Hirsh (2001) uses Latent Semantic Analysis on the combined target and Background Knowledge collections, and observe a substantial increase in classification performance on the target documents. This method assumes that both the target and Background collections are drawn from very similar distributions. In fact for supervised problems it has been shown that this method is often outperformed by the nearest neighbour Bridging algorithm for Background Knowledge of a significantly different structure to the target collection (Zelikovitz (2002)). The method proposed in this paper requires

[2] Due to space constraints a more complete description of the supervised nearest neighbour Bridging algorithm is not included. The interested reader is directed to the literature (Zelikovitz & Hirsh (2002))

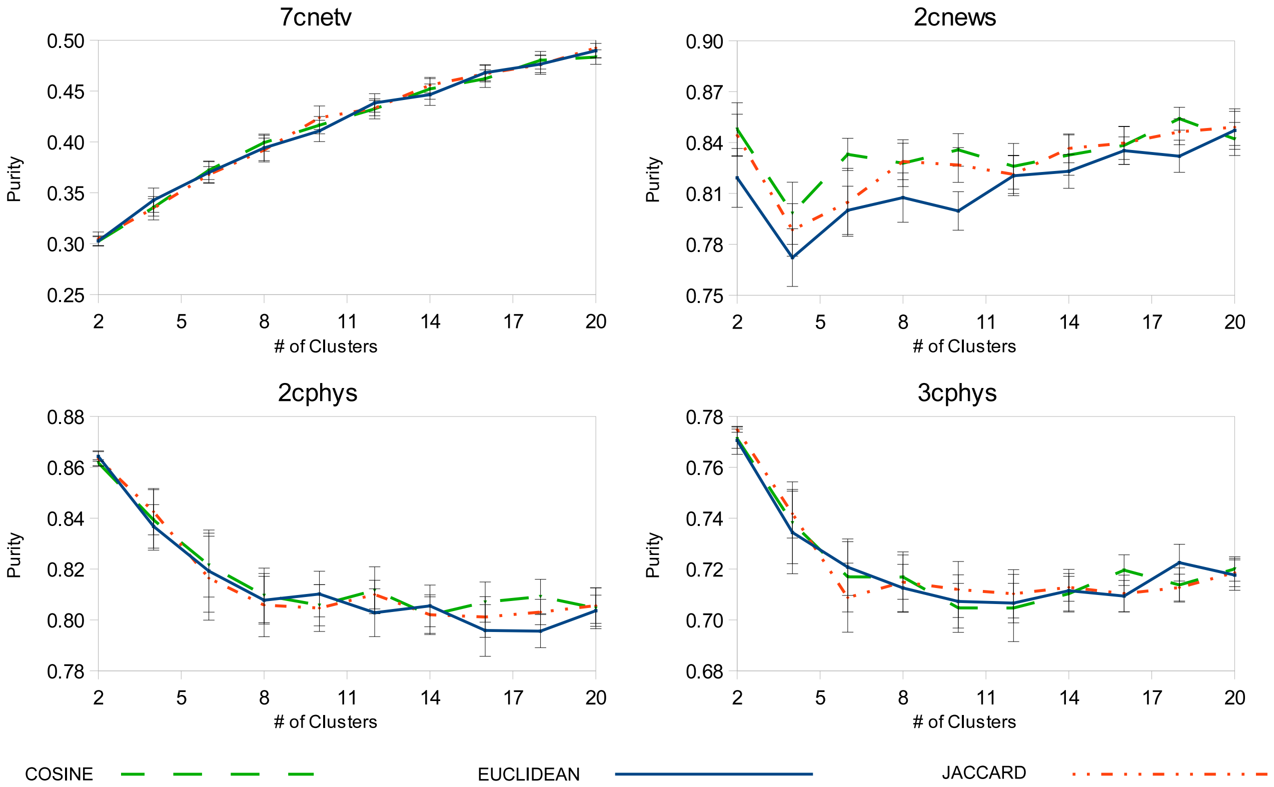

Figure 4: Cluster purity (y-axis) versus number of clusters (x-axis) using the cosine, extended jaccard and euclidean similarity functions in a Bag-of-words space. Graphs are ordered from top to bottom, left to right and present results for the 7CNetv, 2CNews, 2CPhys and 3CPhys datasets respectively.

no such assumption on the Background Knowledge.

5 Experimental Evaluation

In order to evaluate both the Bridging space and the proposed function in terms of their ability to measure similarity between snippets of text, we compare the performance of a number of unsupervised text categorisation tasks both with and without the proposed method. In order to minimise the risk that a negative result is due to a poor choice of Background Knowledge, we perform the evaluation using short text collections for which there is an existing set of Background documents available from the supervised applications of the Bridging space.

The remainder of this section is divided into three parts; a presentation of the datasets employed in our evaluation, a description of the algorithms used to produce the clusters, both in the standard bag-of-words and the Bridging space, and a discussion of the results obtained using our approach.

5.1 Datasets

We now provide a description of each dataset used in this paper. A summary is provided in Table 1. All datasets employed were originally used with Background Knowledge by Zekikovitz et al. (see Zelikovitz (2002)) and are freely available for download[3].

5.1.1 2CNews

The first dataset used is the 2CNews collection, which comprises 1033 news article headings originally published on the ClariNet news site. Each article is labelled as relating to either Business or Sports. The Background Knowledge used is a collection of 1165 partial excerpts from related ClariNet articles that were not included in the test collection.

5.1.2 2CPhys

The second dataset used is a collection of physics technical papers titles. The 2CPhys dataset has a total of 953 titles which are labelled as being related to either astrophysics or condensed matter physics. The Background Knowledge used is 1531 abstracts from other related technical papers.

The 2CPhys dataset provides a more technical problem domain in which to evaluate our proposed approach. The difference in distributions over the classes in 2CPhys is likely to be much more subtle that that for the Business and Sports classes in the 2CNews collection. We expect that this should provide a challenging and interesting task on which to perform our evaluation.

5.1.3 3CPhys

The third dataset used is very similar to 2CPhys in that is is also a collection of titles from physics technical papers. 3CPhys differs however in that it contains 1066 titles, and there are three possible classes (astrophysics, condensed matter physics, and quantum cosmology). The Background Knowledge used with these documents is 1702 abstracts from other related technical papers.

[3] http://www.cs.csi.cuny.edu/~zelikovi/data.htm

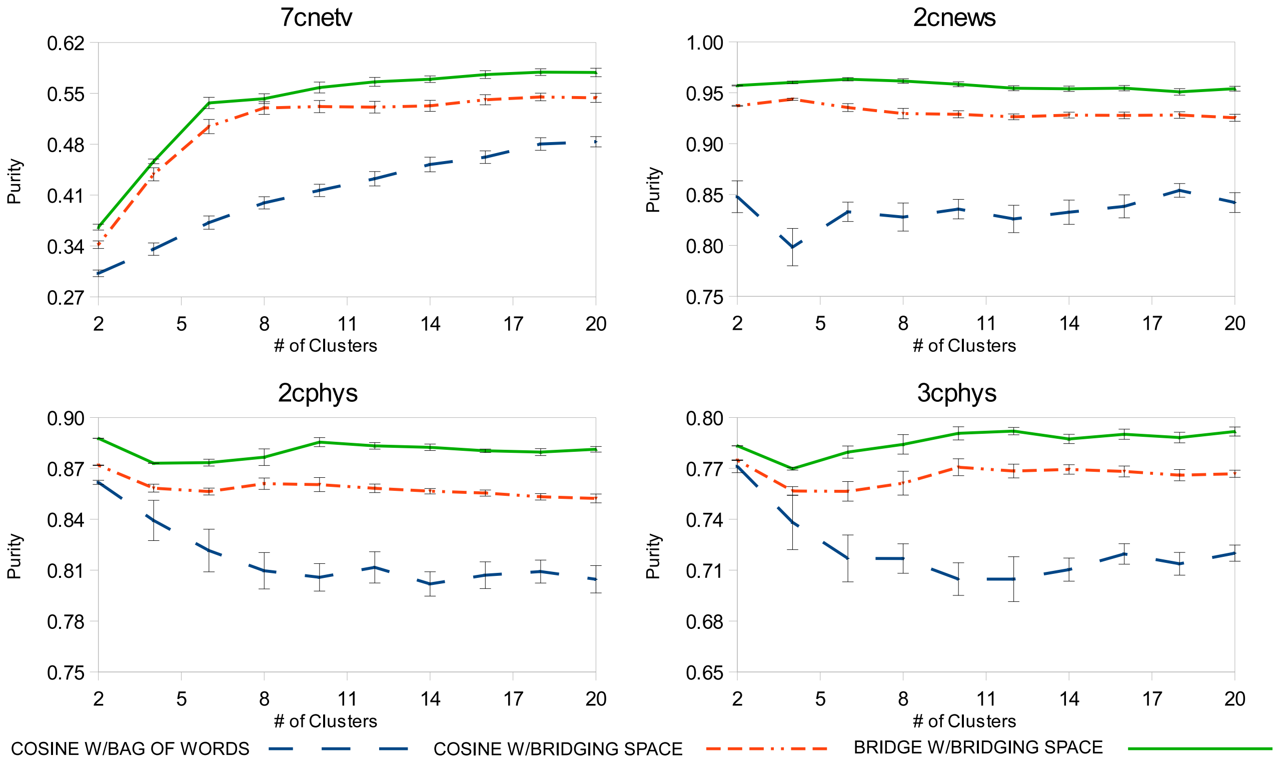

Figure 5: Cluster purity (y-axis) versus number of clusters (x-axis) for the cosine and Bridging functions in the Bridging space. A baseline computed using the cosine similarity in a Bag-of-Words space is also shown. Graphs are ordered from top to bottom, left to right and present results for the 7CNetv, 2CNews, 2CPhys and 3CPhys datasets respectively.

5.1.4 7CNetv

The fourth dataset we use is 7CNetv, web page headings collected from the NetVet website [4]. Each heading is labelled as relating to one of 7 classes; dogs, cats, cows, horses, rodents, primates, and birds. Background Knowledge is created using part of the text from other pages in the NetVet domain. In total there are 1723 test documents and 1160 Background documents. Unlike 2CNews, 2CPhys and 3CPhys the greater number of classes in 7CNetv presents a more complex learning challenge. 7CNetv is also interesting in that the number of Background documents is significantly less than the number of headings.

5.2 Methodology

Document clustering is performed using the freely available CLUTO clustering toolkit (Karypis (2006)). CLUTO is specially designed for use with high dimensional data, and has been used a number of times throughout the text clustering literature (Banerjee et al. (2007), Doucet & Lehtonen (2007), Wang et al. (2008)).

Datasets are input to CLUTO as a matrix of similarity values. In order to construct a similarity matrix for distance functions such as the euclidean distance, we use a value inversely proportional to the equation described in Figure 1, then scale the matrix over the range [0,1]. An objective function f is then selected, along with an algorithm which then attempts to produce the clustering that optimises f over the dataset.

All experiments in this paper were performed using CLUTO's direct clustering algorithm with 10 trial runs. We employed the default objective function which is based on maximising the intra-cluster similarities between instances. All other parameters are left as default settings.

We evaluate the quality of the clustering using the cluster purity evaluation metric (Li et al. (2008), Hu et al. (2008)). All results are averaged over 20 runs and are reported with their 95% confidence interval.

We do not explicitly evaluate any stopping criteria to determine an ideal number of clusters. Instead, for each experiment we vary the number of clusters produced by CLUTO from 2 to 20 (using a step size of 2) and report the purity values over this range of cluster numbers.

The first experiment described in this paper aims to demonstrate that small Background Knowledge collections can be used to significantly increase clustering performance. To the best of our knowledge there have been no previous uses of the datasets described in section 5.1 in the clustering literature. We evaluate the hypothesis by comparing the clustering obtained using the Background Knowledge against a baseline clustering using standard similarity measures in a bag-of-words feature space. Figure 4 shows the purity values obtained using CLUTO with a range of similarity functions and number of output clusters. We observe that the cosine similarity function performs as well or better than all other measures tested with this data. As such we use the cosine similarity function to compute the baseline.

5.3 Results

Figure 5 shows the cluster purity for clustering using the cosine similarity function in both the Bag-of-Words and Bridging spaces for each of the 2CNews, 2CPhys, 3CPhys and 7CNetv datasets. For 2CNews,

[4] http://netvet.wustl.edu

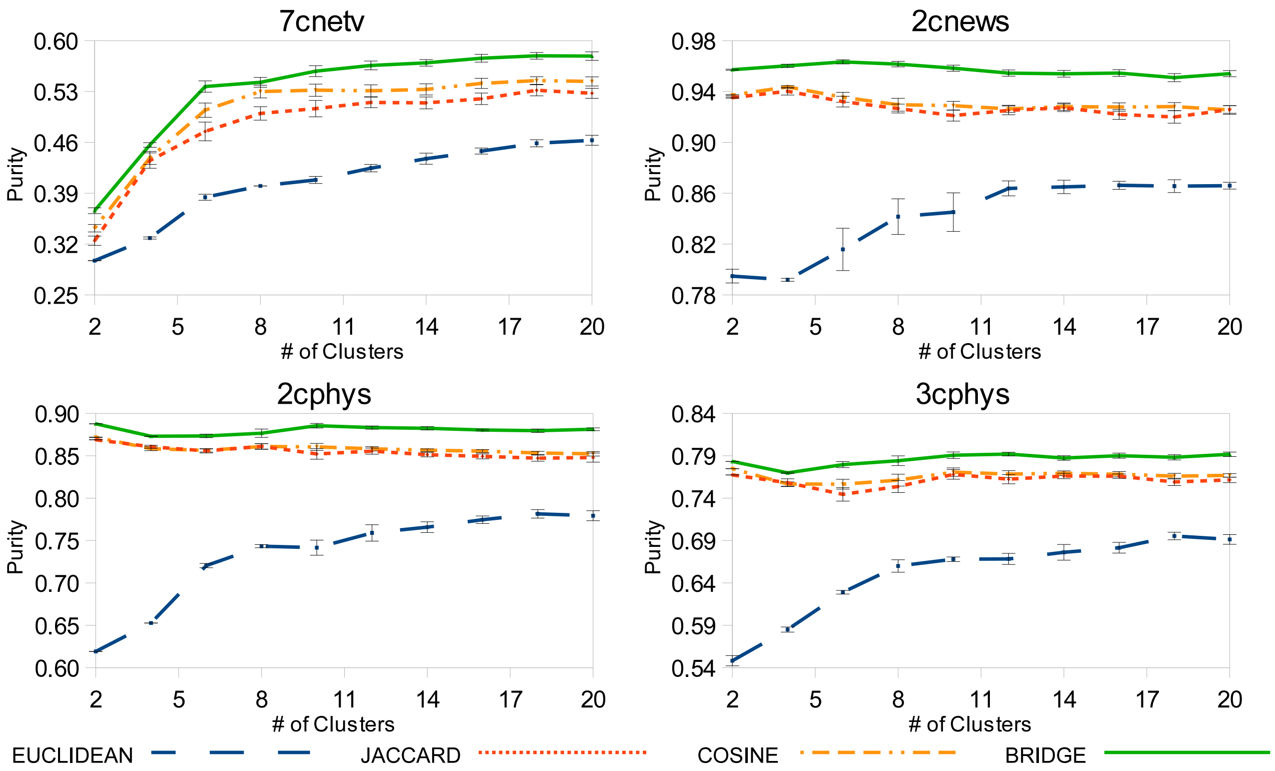

Figure 6: Cluster purity (y-axis) versus number of clusters (x-axis) using the euclidean, extended jaccard, cosine and bridging functions in a Bridging space. Graphs are ordered from top to bottom, left to right and present results for the 7CNetv, 2CNews, 2CPhys and 3CPhys datasets respectively.

7CNetv, and 2CPhys a substantial increase in purity values is observed for all numbers of clusters obtained when using the Background Knowledge to map the target collections into the Bridging space. The purity for the 3CPhys dataset in the Bridging space is also substantially higher than that of the Bag-of-Words representation space for all numbers of clusters tested, with the exception of 2 and 4. We believe this presents a strong case that using Background Knowledge to reexpress text in the Bridging space can improve the ability of clustering algorithms to measure similarity between short text documents.

We note that when clustering the 2CNews, 2CPhys and 3CPhys in the Bridging space measured purity values are relatively consistent as the number of clusters is varied. However with 7CNetv a very sharp increase in purity is observed as the number of clusters is increased from 2 to 6. We hypothesise that this is due to the number of hidden classes in 7CNetv being larger than the number of clusters produced. For example when generating 4 clusters for the 7CNetv dataset, all documents from at least 3 hidden classes will be counted as being incorrectly clustered.

Figure 6 compares the measured purity values obtained using the proposed clustering function to that of the cosine, euclidean, and jaccard functions (see figure 1) when clustering documents in the Bridging space. For all datasets and number of clusters evaluated, we note that the proposed clustering function substantially outperforms all other functions tested. This demonstrates the effectiveness of the proposed function when clustering documents that have been represented in the Bridging space.

We note that when clustering using the proposed function in the Bridging space, the measured purity values for all datasets are consistently greater than those for the best baseline (see figure 5). This provides a strong case for the use of the proposed function on data represented in the Bridging space.

We note the relatively poor performance of clustering using the euclidean distance compared to the cosine, jaccard, and proposed Bridging functions. As the euclidean distance is based on the difference of individual attribute values, euclidean distance will not distinguish between features with identical, high values for each vector (ie. from Background documents that share terms with both short text strings), and features for which both vectors are zero (ie. Background documents that share no terms with either snippet). This behaviour is not shared by the other comparison functions tested in this work, which are based on the product of attribute values and therefore ignore features for which both vectors are zero. Recall also that as the cosine and extended jaccard features are sensitive to the length of the vectors that they will be effected by the features for which only one of the vectors is non-zero. The relative performances of the cosine, euclidean, jaccard, and Bridging functions supports proposition 1 given in section 3; namely that Background documents sharing no terms with one or both of the short text strings should not influence the result of the function.

6 Conclusions

In this paper we have presented a method for leveraging relatively small, unlabelled collections of semantically relevant documents to improve the clustering of short text data. We refer to this unlabelled, relevant text as Background Knowledge. A summary of the novelty and major contributions of this paper is as follows:

- We demonstrate a simple method for using Background Knowledge to construct an alternative representation for short text called the Bridging space. We show that using Background Knowledge with this method significantly increases cluster purity. While the Bridging space has been used before in supervised document categorisation, to the best of our knowledge this is the first such use for unsupervised clustering.

- Unlike much of the previous literature concerning external document collections, the Background Knowledge corpora we employ are small and contain only a few thousand documents each. The use of simple methods to exploit small Background Knowledge collections is novel. The reduction in the size of the external collections is likely to provide significant practical benefits with regards to obtaining and use of the external corpora.
- We propose a clustering function for use on short text documents represented in the Bridging space. Experimental results have shown this method to be very effective, outperforming standard distance and similarity measures by a substantial margin on four separate document collections.

A number of directions for future work exist. Possible extensions to the work described in this paper include:

- The value of k used in equation (2) was 10 for all experiments reported in this paper. While we have achieved good results with this value setting a more formal evaluation of the ideal value would be of interest.
- Exactly what makes an effective corpus of Background Knowledge is dependent on the specific clustering task at hand. For a given text clustering problem, finding an appropriate set of Background Knowledge is by no means a trivial task. While there has been research in the supervised literature on automatically obtaining Background Knowledge for a given dataset (Zelikovitz & Kogan (2006)), the methods described are specific to supervised learning. Exploring methods for obtaining Background Knowledge for use with unsupervised tasks would be useful.
- The application of Background Knowledge along with the proposed similarity function has been shown to significantly increase the purity when clustering short text documents. An explicit comparison of small and large external document corpora was however not performed. A comparison of small collections of Background Knowledge with alternative sources of external knowledge such as Wikipedia would be interesting.

References

Banerjee, S., Ramanathan, K. & Gupta, A. (2007), Clustering short texts using wikipedia, *in* 'SIGIR'.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research* **3**, 993–1022.

Blum, A. & Mitchell, T. (1998), Combining labeled and unlabeled data with co-training, *in* 'COLT: Proceedings of the Workshop on Computational Learning Theory', pp. 92–100.

Bollegala, D., Matsuo, Y. & Ishizuka, M. (2007), Measuring semantic similarity between words using web search engines, *in* 'WWW'.

Chan, J., Koprinska, I. & Poon, J. (2006), Nearest neighbour classification with background knowledge extended to semi-supervised learning, Technical report, University of Sydney.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A. (1990), 'Indexing by latent semantic analysis', *Journal of the American Society of Information Science* **41**(6), 391–407.

Doucet, A. & Lehtonen, M. (2007), Unsupervised classification of text-centric xml document collections, *in* 'INEX'.

Gabrilovich, E. & Markovitch, S. (2006), Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge, *in* 'AAAI'.

Gabrilovich, E. & Markovitch, S. (2007), Computing semantic relatedness using wikipedia-based explicit semantic analysis, *in* 'IJCAI'.

Gupta, R. & Ratinov, L. (2008), Text categorization with knowledge transfer from heterogeneous data sources, *in* 'AAAI'.

Hofmann, T. (1999), Probabilistic latent semantic analysis, *in* 'Proc. of Uncertainty in Artificial Intelligence, UAI'99', Stockholm.

Hotho, A., Staab, S. & Stumme, G. (2003), Wordnet improves text document clustering, *in* 'In Proceedings of the SIGIR Semantic Web Workshop'.

Hu, X., Sun, N., Zhang, C. & Chua, T. (2008), Enhancing text clustering by leveraging wikipedia semantics, *in* 'SIGIR'.

Hu, X., Sun, N., Zhang, C. & Chua, T. (2009), Exploiting internal and external semantics for the clustering of short texts using world knowledge, *in* 'CIKM'.

Jing, L., Zhou, L., Ng, M. & Huang, J. (2006), Ontology-based distance measure for text clustering, *in* 'SIAM International Conference on Data Mining'.

Joydeep, A. S., Strehl, E., Ghosh, J. & Mooney, R. (2000), Impact of similarity measures on web-page clustering, *in* 'In Workshop on Artificial Intelligence for Web Search (AAAI 2000)', AAAI, pp. 58–64.

Karypis, G. (2006), Cluto-a clustering toolkit, release 2.1.1, Technical report, Department of Computer Science, University of Minnesota. http://www.cs.umn.edu/ karypis/cluto/.

Li, Y., Chung, S. & Holt, J. (2008), Text document clustering based in frequent word meaning sequences, *in* 'Data and Knowledge Engineering'.

Miller, G. A. (1995), 'Wordnet: A lexical database for english', *Communications of the ACM* **38**(1), 39–41.

Phan, X., Nguyen, L. & Horiguchi, S. (2008), Learning to classify short and sparse text and web with hidden topics from large-scale data collections, *in* 'WWW'.

Sahami, M. & Heilman, T. (2006), A web-based kernel function for measuring the similarity of short text snippets, *in* 'WWW'.

Tversky, A. (1977), 'Features of similarity', *Psychological Review* **84**(4), 327–352.

Wagstaff, K., Cardie, C., Rogers, S. & Schrodl, S. (2001), Constrained k-means clustering with background knowledge, *in* 'ICML', pp. 577–584.

Wang, P., Domeniconi, C. & Hu, J. (2008), Using wikipedia for co-clustering based cross-domain text classification, *in* 'ICDM'.

Weng, G. & Poon, J. (2006), A data complexity analysis on imbalanced datasets and an alternative imbalance recovering strategy, *in* 'WI'.

Yih, W. & Meek, C. (2007), Improving similarity measures for short segments of text, *in* 'AAAI'.

Zelikovitz, S. (2002), Using Background Knowledge to Improve Text Classification, PhD thesis, Rutgers University.

Zelikovitz, S. & Hirsh, H. (2001), Using lsi for text classification in the presence of background knowledge, *in* 'CIKM'.

Zelikovitz, S. & Hirsh, H. (2002), Integrating background knowledge into nearest-neighbor text classification, *in* 'ECCBR'.

Zelikovitz, S. & Hirsh, H. (2005), Improving text classification using em with background text, *in* 'FLAIRS'.

Zelikovitz, S. & Kogan, M. (2006), Using web searches on important words to create background sets for lsi classification, *in* 'FLAIRS'.