Selective Integration of Background Knowledge in TCBR Systems

Anil Patelia[1], Sutanu Chakraborti[1], and Nirmalie Wiratunga[2]

[1]Department of Computer Science and Engineering,
Indian Institute of Technology Madras, Chennai–600036, India
[2]School of Computing, The Robert Gordon University,
Aberdeen AB25 1HG, Scotland, UK
pateliaaj@cse.iitm.ac.in, sutanuc@iitm.ac.in, n.wiratunga@rgu.ac.uk

Abstract. This paper explores how background knowledge from freely available web resources can be utilised for Textual Case Based Reasoning. The work reported here extends the existing Explicit Semantic Analysis approach to representation, where textual content is represented using concepts with correspondence to Wikipedia articles. We present approaches to identify Wikipedia pages that are likely to contribute to the effectiveness of text classification tasks. We also study the effect of modelling semantic similarity between concepts (amounting to Wikipedia articles) empirically. We conclude with the observation that integrating background knowledge from resources like Wikipedia into TCBR tasks holds a lot of promise as it can improve system effectiveness even without elaborate manual knowledge engineering. Significant performance gains are obtained using a very small number of features that have very strong correspondence to how humans describe the domain.

1 Introduction

Textual Case Based Reasoning (TCBR) aims at solving new problems by reusing past experiences recorded in the form of free form (or semi-structured) text. The effectiveness of TCBR systems is critically dependent on the method used to estimate semantic relatedness between two pieces of text. As humans, we are skilled at arriving at representations that capture deeper meanings of texts that may not have a direct bearing with the surface level word forms. In doing so, we not only use an elaborate knowledge of language, but also implicitly and seamlessly integrate common-sense and background knowledge. It is thus natural to suppose that TCBR systems would also benefit from a principled integration of background knowledge. This paper reports experiments we conducted towards testing this hypothesis. A comparative study on text classification shows that background knowledge as is readily available in resources like Wikipedia can lead to improvements in retrieval effectiveness.

We extend the Explicit Semantic Analysis (ESA) [2] approach to allow easy integration of Wikipedia knowledge into instance based learners. The key idea is to treat Wikipedia articles as concepts and construct representation of documents

A. Ram and N. Wiratunga (Eds.): ICCBR 2011, LNAI 6880, pp. 196–210, 2011.

as feature vectors over these concepts. Intuitively, the relevance of a concept to a document is estimated by measuring the overlap of the words present in the document and those present in the Wikipedia article corresponding to the concept. This approach lends itself easily to a TCBR framework since it allows for lazy incremental learning that relies on local models. Also, true to the spirit of CBR, the ESA representations are easily interpretable and retrieval or classification results can be easily explained. There are two significant questions that remain unanswered: how do we identify the set of Wikipedia articles (concepts) relevant to a given task? and can we do better by relaxing the assumption that the Wikipedia concepts are unrelated to each other? In other words, can we enrich the retrieval performance of the system by modelling the relatedness of Wikipedia concepts?

The paper is organized into the following sections. Section 2 presents a background to our work and identifies related works. Section 3 introduces four different Wikipedia article selection strategies, and section 4 describes an approach to estimate semantic relatedness between Wikipedia articles, and integrate this knowledge to obtain revised representation of cases. Section 5 presents empirical evaluation of our approaches.In Section 6, we deliberate on the key ideas behind this paper and reflect on certain research directions that are motivated by this work. Section 7 summarizes our key contributions.

2 Background and Related Work

Let us consider an example to motivate the importance of background knowledge in estimating relatedness of documents. Considering two short documents describing chess moves, one containing the word "rook" and another containing the word "bishop". If these documents share no other term, a TCBR system may not be able to relate the two documents. However, the two words rook and bishop co-occur in the Wikipedia article on chess. In this way Wikipedia knowledge can help in arriving at better models of semantic similarity between words, as well as between documents. The key idea behind ESA [2] is to treat words (and phrases) like chess as general concepts and express documents (textual cases) in terms of these concepts. More specifically, each Wikipedia article is thought of as representing a concept and each document, as well as each word, is a vector over a space defined by these concepts.

Figure 1 shows an example to illustrate the idea behind ESA. The sentences "US President summarizes his position on the Middle East" and "Israel and Palestine respond to Obamas foreign policy note" share no words. Yet we know that these sentences are strongly related to each other. The Wikipedia articles to which the words in sentence 1 has strong correspondence include {Barack_Obama, Foreign_policy_of_the_United_States, Middle_East, Afganistan}, which has a good overlap with the set of Wikipedia articles { Barack_Obama, Middle_East, Foreign_policy_of_the_United_States } that are related to sentence 2. The two sentences were orthogonal to each other in the original vector space spanned by words, but display high similarity when represented in the new vector space, where each dimension corresponds to a concept which maps on to a

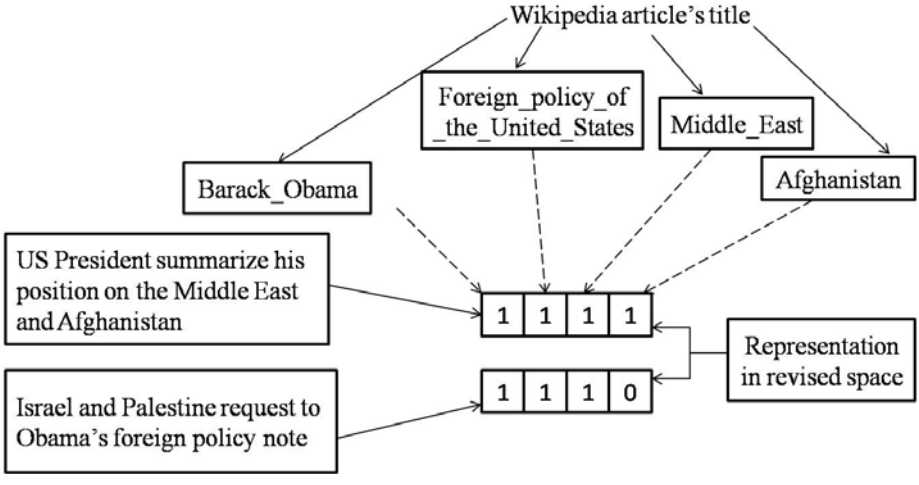

Fig. 1. Representation of news sentences in revised space

Wikipedia article name. For a given domain, we would like to restrict attention to a small set of relevant concepts.

Whilst concepts can be constructed introspectively (e.g. Latent Semantic Indexing) it is only possible when concepts needed to construct rich representations in a domain are present within the document descriptions themselves. Consider the two sentences in Figure 1, which can only be related if we know something about politics in the US and in the Middle East. ESA has access to the world that defines the context, and can overcome this limitation. Also, concepts are Wikipedia article names, which humans find easy to relate to. Accordingly ESA provides an elegant means to incorporate background knowledge in a transparent manner.

There have been much research aimed at creating revised representations of documents based on linguistic or background knowledge. Scott et al [11] used the synonymy and hypernymy relations from WordNet[3] to revise bag-of-words representations. Zelikovitz et al. [10] present a case for transductive learning, whereby test documents (without their class labels) were treated as a source of background knowledge to make up for the inadequacy of labeled examples. Others have also attempted to mine relationships between entities in Wikipedia. This is useful for tasks like constructing domain specific resources like thesauri, taxonomies and ontology. In the CBR community, Propositional Semantic Indexing [7] has been proposed as an alternative to approaches like LSA [5]. PSI features are more expressive than those derived from LSI in that they are logical combinations of words (as opposed to linear combinations in LSI), however they can only be composed out of existing words. This means that a compact concept descriptor like, chess, or US Politics, is highly descriptive of how we view the domain cannot emerge as new features. An extracted feature in PSI can at best be a disjunction over conjunctions of several terms related to chess, and this may

become unwieldy and hard to understand as the descriptions grow longer. PSI is an introspective learner and has no access to background knowledge.

2.1 Using ESA for Classification

Figure 2 illustrates how ESA can be used to represent cases for text classification tasks. Each training document is mapped to a concept representation. A concept corresponds to a Wikipedia article. The semantic similarity of each term to a concept is estimated by observing how strongly (say in terms of a tf-idf measure) the term is present in a Wikipedia article corresponding to that concept. Once we have representation of each term as a concept vector, a document (case) can be represented as a concept vector as well. The concept vector representing the document is simply the vector sum of the concept vectors corresponding to each term present in the document. The unseen test document is mapped to its concept representation, which is compared against the concept vectors of training documents, and the top k training documents according to a cosine similarity measure are used to decide the class label of the test document.

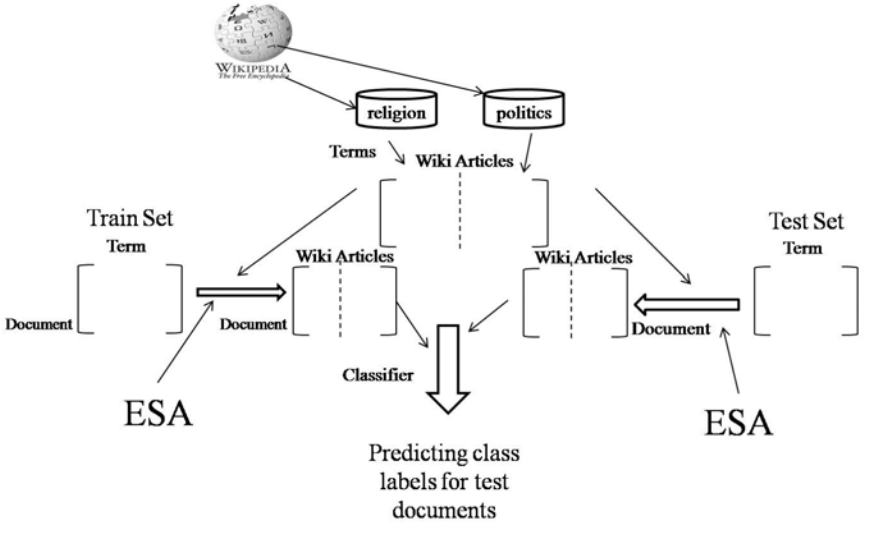

Fig. 2. Explicit Semantic Analysis for classification

3 Informed Selection Strategies for Wikipedia Articles

While incorporation of background knowledge from Wikipedia can be useful in improving system effectiveness, it is also important to know which Wikipedia pages to actually use for modeling concepts, given a specific task like text classification and a corpus of documents (cases). One option is to look at all Wikipedia articles that contain any of the terms used in the training corpus. This may result

in accumulating web-pages that are also remotely relevant to the classification task. Interestingly, Wikipedia pages are tagged with knowledge of categories (drawn from a hierarchy), and this can act as a preliminary filter. For example, in a text classification scenario where we want to discriminate between documents of classes Religion and Politics, we may only consider Wikipedia pages belonging to those categories. Sometimes, the category labels in Wikipedia will not have neat correspondence to the class labels of the domains, but it is often possible to establish a mapping. This approach of considering all pages under certain Wikipedia categories is often not adequate. Since not all Wikipedia pages tagged with the relevant category labels will help in discriminating between the classes. Thus, we may still have a large number of redundant Wikipedia pages being considered. We proposed and experimented with four different Wikipedia article selection strategies with the goal of addressing these shortcomings.

Baseline Approach. The Baseline algorithm used for our comparisons is one that compiles a collection of Wikipedia articles that have category labels relevant to the classification task, and randomly selects pages from this collection to generate ESA representations. Therefore the baseline algorithm is top-down and solely driven by category labels in the collection. Essentially it completely disregards bottom-up clues from the words actually used in the training corpus.

3.1 Centroid Strategy

The centroid strategy is founded on a vector space that is spanned by the union of all distinct words in the domain, and those that appear in Wikipedia articles relevant to the class labels. For each class, we compute the centroid of training documents in that class. Wikipedia articles are ranked based on maximum cosine similarity they have with any cluster centroid, and the top k articles are selected. Figure 3 illustrates the idea and summarizes the algorithm. The basic idea behind this approach is to select web-pages prototypical of the

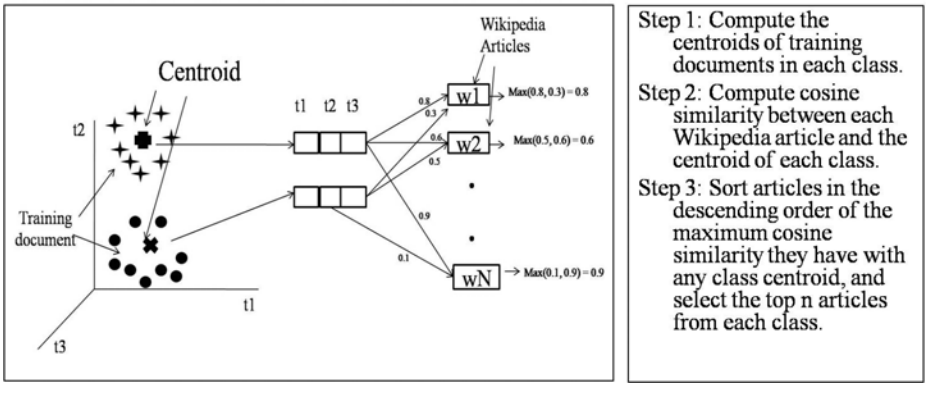

Fig. 3. Centroid strategy for Wikipedia article selection

categories. However, one downside of this approach is that we could imagine pathological situations where certain categories starve. In other words, we are not guaranteed to obtain adequate number of representative Wikipedia pages for each category. The second limitation is that a Wikipedia article could be very prototypical of more than one class, in which case it may not be very good at discriminating between classes, even if it is ranked highly. A third limitation arises from the observation that there may be scenarios where the cluster centroids are not adequately representative of the Wikipedia pages in the corresponding categories. This situation is common in complex classification tasks where Wikipedia pages in disjoint well separated clusters are labelled with the same category tag.

3.2 k-Nearest Neighbour Strategy

In this approach, we no longer use the centroid as a representative of a class. Instead, corresponding to each Wikipedia article we identify the training documents that are closest to it in terms of the cosine similarity. A rank is assigned to a Wikipedia article based on the sum of the top three cosine similarities. The top ranked Wikipedia articles are treated as concepts for classification. Figure 4 illustrates the idea and summarizes the algorithm. This approach overcomes a key limitation of the centroid approach, in that it can handle complex classification problems where local neighbourhoods are more indicative of correct category than proximity to class centroid. A limitation of this could be that Wikipedia pages that are extremely similar to each other can get selected, leading to redundancy.

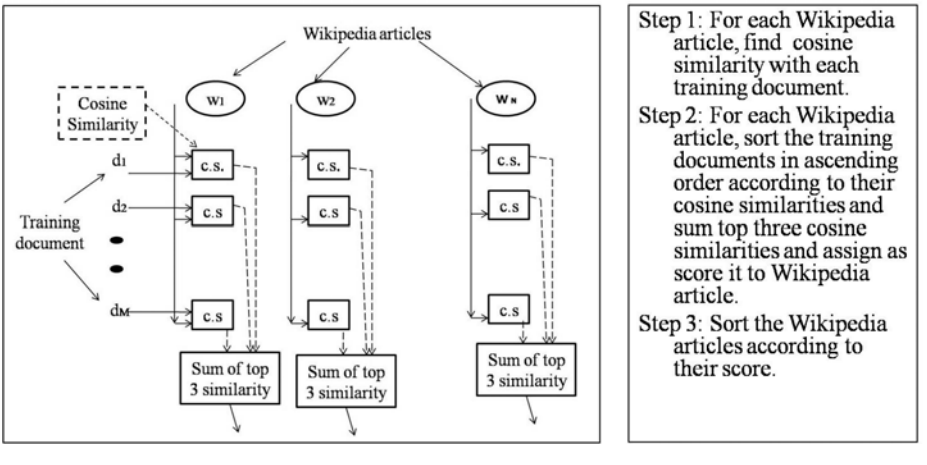

Fig. 4. kNN strategy for Wikipedia article selection

3.3 k-Nearest Neighbour with Discrimination Strategy

This strategy is very similar to the kNN approach, except that we ensure that each class is assigned representative articles. Thus we select the top m Wikipedia articles for each class having highest cosine similarities with the three nearest neighbours in training documents of that class. This overcomes the second limitation of the Centroid strategy in that it guarantees that no class suffers from starvation.

3.4 Probability Ratio Strategy

This strategy evaluates the relative importance of a Wikipedia article to a class using probability estimates computed from the training corpus using add-1 smoothing. Given a class c drawn from a set of n categories, the posterior probability $P(c|wk)$ of c given a Wikipedia article, wk is estimated. A Naive Bayes Classifier that assumes conditional independence of the features is used [9]. The Wikipedia article is assigned to the class which gives rise to the highest posterior estimate. The top few Wikipedia articles of each category are selected.

3.5 Augmented ESA

In addition to the four Wikipedia article selection strategies described above, we also carried out an experiment where a representation of a textual case was formed using a mix of words and concepts derived from Wikipedia. This was motivated by the observation that we do not wish to loose those words that are already good in discriminating between classes. This approach is referred to as Augmented ESA.

We tried an approach that attempts to directly estimate the discriminating power of a Wikipedia page, and selects those pages that allow for best discrimination between classes. A Wikipedia page is represented in terms of a vector of real valued tf-idf values over the feature space of words. So we need a discretization (binning) method to evaluate the Information Gain of the concept feature corresponding to that page. When only two bins are used, we get a binary-valued feature corresponding to each concept. The details of this binning approach are available in [13]. These discretrized concept features are stacked along with binary valued features derived from words, as in Augmented ESA, and the Information Gain of the word-level features as well as those of concept level features are evaluated. The concept features (articles) having highest Information Gain are selected. It may be noted that while using the Information Gain idea, no filtering mechanism is used to prune the set of Wikipedia articles. Rather, a huge number of relevant articles are evaluated for their Information Gain, without consideration of whether they have significant correspondence to the documents in the training corpus.

4 Modelling Similarity between Wikipedia Articles

The ESA approach assumes that each Wikipedia article represents a concept that is unrelated to all other concepts (Wikipedia articles). It is easy to see that this is at best a convenient approximation. In this section, we discuss how we incorporate the knowledge of similarity between Wikipedia articles into the revised representation of terms and documents.

A Case Retrieval Network (CRN) is used to capture the pair wise concept similarities which are used to revise the document representations. Let us consider a document as being represented as a vector, each component of which represents the relevance of the document to a concept. We can assume that these relevances are zero when a concept is not relevant to a document and 1 when it is relevant. In the CRN framework, we have similarity arcs connecting every pair of concepts as shown in Figure 5. The relevant concepts are allowed to "activate" other concepts which have non-zero similarity to it, using a process of spreading activation. At each concept node the incoming activations are aggregated and the revised document representation is a vector compromising the aggregated activation at each concept node. For example assume an initial representation of document D is 1, 1, 0, 0, 0 in the vector space of Wikipedia-based concepts W1 through W5. Let us consider the pair wise similarity values as shown in Figure 5. If the aggregation function at each node is a simple summation, the resulting representation of D ought to be 1.9, 1.9, 1.1, 0, 0. This new representation can be seen as a result of a matrix operation. Let R_i and R_n be initial and new representation of the document respectively and S be a symmetric matrix of concept pair similarities. The new representation can be given as $R_n = R_i$S.

The similarities between Wikipedia articles are estimated using Latent Semantic Analysis (LSA). This is in line with an earlier work where LSA was used for introspective knowledge acquisition in CRNs[1]. In particular we use Sprinkled LSA[8] whereby category knowledge is incorporated into the process of obtaining revised lower dimensional representations. This is motivated by the observation that while LSA dimensions capture significant variances in the data, they are not guaranteed to be the ones with highest discriminatory power. The central idea behind sprinkling is to augment a document representation with additional terms, each representative of a particular category to which the document belongs. This has the effect of pulling together documents belonging to the same category and emphasizing the distinction between documents belonging to

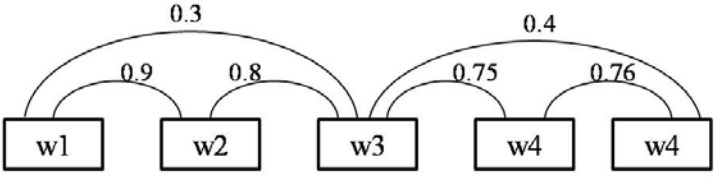

Fig. 5. Case retrieval network

different categories. The number of sprinkled (augmented) terms can be varied to control the degree to which category knowledge is emphasized. The details of this procedure are explained in [8].

5 Evaluation

We evaluated the effectiveness of our proposed integration of background knowledge from Wikipedia in the context of text classification.

5.1 Datasets and Methodology

We tried classification on four datasets created from the 20 Newsgroups [6] corpus. There are a total of twenty different news-group categories in this dataset. Each category has thousand articles drawn from postings of discussions, queries, comments etc. Four datasets were formed from the news-group:

- HARDWARE group from two hardware categories, one on MAC and the other on PC.
- RELPOL, from two groups, one concerning religion, the other politics in the middle-east.
- SCIENCE from four science related groups
- REC from four recreation related groups.

Thus HARDWARE and RELPOL are two class problems, and SCIENCE and REC are multi-class problems. Each sub-corpus was divided into train and test sets. Sizes of train and test sets are equal. Each partition contains 20% of documents randomly selected from the original corpus, and is stratified in that it preserves the class distribution of the original corpus. Fifteen such train-test splits (alternately called trials) were obtained for each of the four datasets mentioned above. It may be noted that the documents were pre-processed by removing stop words (noise words) like functional words which are frequent throughout the collection and ineffective in discriminating between classes. Weighted kNN classifier is used with $k = 3$.

Table 1 compares the accuracies of ESA against a naive bag-of-words Vector Space approach and the Baseline on 4 sub category from the 20Newsgroup. Table 2 reports the accuracies obtained when Augmented ESA representation (see Section 3.5) using a mix of concepts and words were used. As we can see, ESA techniques yield substantial improvements over Vector Space Model in each category. ESA with various Wikipedia article selection strategies also achieves much better accuracy compared to the Baseline approach that relied on an adhoc selection procedure. Results presented in Figures 6 to 9 generally suggests that classification accuracy increases as a function of the number of Wikipedia pages. This is particularly evident RELPOL and HARDWARE, where the increase is steeper than with random selection of Wikipedia articles. This shows that we can attain conspicuous improvements using fewer pages, if we adopt a principled approach to selection of Wikipedia articles.

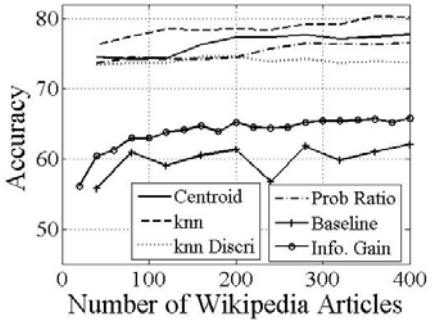

Fig. 6. ESA on HARDWARE

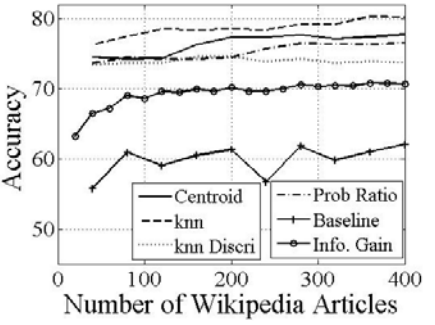

Fig. 7. Augmented ESA on HARDWARE

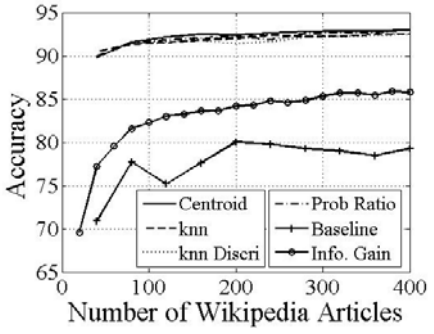

Fig. 8. ESA on RELPOL

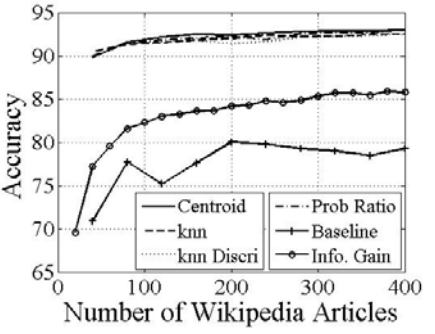

Fig. 9. Augmented ESA on RELPOL

Table 1. Comparison of performance of ESA, Vector Space Model (VSM) and Baseline

Dataset	Wikipedia document Selection strategy				Info gain	VSM	Baseline
	Centroid	Knn	knnDiscri	probRatio			
HARDWARE	77.72	76.51	74.60	**80.28**	65.79	59.51	65.77
RELPOL	92.94	92.98	92.43	92.54	85.76	70.51	80.88
SCIENCE	**81.01**	76.76	78.34	77.98	68.90	54.89	60.22
RECREATION	**83.02**	76.76	79.72	77.26	67.99	62.79	66.54

Table 2. Performance of Augmented ESA

Dataset	Wikipedia document Selection strategy				Info gain	VSM	Baseline
	Centroid	Knn	knnDiscri	probRatio			
HARDWARE	74.75	76.70	76.51	75.84	70.69	59.51	65.77
RELPOL	93.13	93.09	93.04	93.09	85.93	70.51	80.88
SCIENCE	77.76	78.16	76.44	77.63	71.88	54.89	60.22
RECREATION	**82.68**	77.45	77.31	79.23	70.32	62.79	66.54

Table 3. Performance of ESA with knowledge of concept similarities

Dataset	Wikipedia document Selection strategy				Info gain	VSM	Baseline
	Centroid	Knn	knnDiscri	probRatio			
HARDWARE	75.37	75.37	72.38	**78.12**	69.98	59.51	65.77
RELPOL	**93.08**	92.64	91.04	92.49	86.45	70.51	80.88
SCIENCE	**79.69**	77.91	76.13	75.00	70.56	54.89	60.22
RECREATION	**80.05**	72.89	76.43	75.66	70.37	62.79	66.54

Table 4. Performance of Augmented ESA with knowledge of concept similarities

Dataset	Wikipedia document Selection strategy				Info gain	VSM	Baseline
	Centroid	Knn	knnDiscri	probRatio			
HARDWARE	**76.57**	74.23	74.23	75.77	72.56	59.51	65.77
RELPOL	94.49	94.31	93.66	94.23	86.88	70.51	80.88
SCIENCE	**82.49**	80.83	79.95	79.78	72.38	54.89	60.22
RECREATION	**80.87**	78.56	77.82	79.89	73.21	62.79	66.54

5.2 Modeling Similarity between Wikipedia Articles

We empirically evaluated the impact of modelling similarity between Wikipedia pages as described in Section 4. We use Latent Semantic Indexing for modelling similarity between Wikipedia articles. The main parameter in LSI is the number of dimensions used, which should ideally be set using cross validation. The results reported in this section correspond to choice of dimensions that led to best LSI performances. Table 3 shows the classification accuracy using the revised case representation which incorporates knowledge of similarities between concepts. Table 4 shows the results when Augmented ESA representation is used, along with knowledge of similarities between concepts.

5.3 Summary of Observations

Paired one tailed t-test with 95% confidence was used to analyse the observed differences between accuracies reported by each pair of methods over the 15 train test pairs. We observe that after integration of background knowledge, effectiveness of text classification improves conspicuously. The improvements are more pronounced when the principled article selection strategies described in Section 3 are used. Importantly significant gains are seen even when fewer concept-level features are used. As shown in Table 1, Baseline algorithm performs significantly better than naive Vector Space model for each category and differences vary from 3% to 10%. Each Wikipedia article selection strategy performs significantly better than the baseline. The classification accuracy of RELPOL dataset increases from around 80% in Baseline to more than 90% for each of the different Wikipedia article selection strategies as shown in Table 1.

Comparing the Wikipedia article selection strategies, we can see that the centroid strategy is, on the whole, better than the rest. As shown in Table 2, Augmented ESA performs better than ESA on RELPOL dataset. Similarity modelling between Wikipedia articles does not improve the result for ESA presented in Table 3. In particular, the highest accuracy is 94.49% when centroid strategy is used for augmented ESA representation with similarity modelling. Augmented ESA with similarity modelling performs better than case representations that ignore original features for SCIENCE dataset, over all Wikipedia articles selection strategies. For the HARDWARE dataset, performance decreases as we try to model similarity between Wikipedia articles. A closer look suggests that there are many common Wikipedia articles belonging to both categories. For instance we have observed that articles like persona_computer, history_of_computing_hardware are selected for both categories of hardware (IBM and Mac). These pages seem to be related to both hardware.ibm and hardware.apple, and hence cannot help in the classification task. Justification for additional similarity modelling as discussed in Section 4 remains weak. For instance we found that there was a significant difference between the similarity modelling versions of ESA (except on the RELPOL dataset), where augmented ESA was found to be better.

It is interesting to note from Tables 1 through 4 that the four principal Wikipedia article selection strategies described in Section 3 far outperform the Information Gain based measure outlined in Section 3.5. This can be attributed to the fact that the Information Gain measure ignores the bottom-up information suggested by the actual words in the training corpus. The article selection strategies appear to strike a decent trade-off between selecting features that are inspired by the corpus, and those that actually contribute positively to discrimination between classes. Also, Figures 6 through 9 show that all four article selection strategies lead to performance improvements even with fewer Wikipedia articles (of the order of 50 to 100). The Information Gain based measure is less robust and shows a sharper increase as more Wikipedia based concepts are included. Improvements with injecting semantic similarity between concepts was not very pronounced, except in a few domains over select article selection strategies. In retrospect, we perhaps need to be more conservative in linking up Wikipedia articles. We may like to add measures of similarity after evaluating their potential impact on classification accuracies. The semantic similarity computation may also need to be refined by incorporating knowledge of hyperlink associations and category links attached to Wikipedia pages.

The improvement gains over complex datasets like Hardware are really encouraging. It may be noted that Apple and Mac classes in HARDWARE have good overlap of terms they share, but the classes get more easily separable when background knowledge is in place. The kNN strategy outperforms the rest in Hardware, whereas in RELPOL the differences between strategies are less pronounced. This hints at the fact that local models work better in complex domains as opposed to global ones (like centroid based strategies), as they lend themselves to modeling more complex decision boundaries.

6 Discussion and Outlook

An interesting aspect of the integration of background knowledge using principled Wikipedia article selection strategies is the fact that we can achieve significant gains in effectiveness using very few features. The graphs in the previous section illustrate that knowledge of around 30-40 semantically rich concepts constituting background knowledge in the domain is perhaps more worth knowing than blindly acquiring thousands of word-level features with the hope of learning statistical models that are severely constrained by the representativeness of the data they are presented with, and are hard to train, interpret and maintain. It is important to know which concepts will make the most impact given the task and the dataset at hand; article selection strategies presented in this paper are designed with this goal in mind.

Having very few features has implications in terms of improving retrieval efficiency. While efficiency and effectiveness are often viewed as conflicting goals, it turns out that having fewer concepts can contribute positively to realising both these goals at the same time. Hubert Dreyfus observes: AI researchers have long recognized that the more a system knows about a particular state of affairs, the longer it takes to retrieve the relevant information, and this presents a general problem when scaling up is concerned. Conversely, the more a human being knows about a situation or an individual, the easier it is to retrieve other relevant information. Additionally having access to a knowledge repository as large as the WWW can help CBR systems do better than just look at the set of cases it has immediate access to. Several research strategies view the WWW as a means to provide access to many more cases. This view can be restrictive in that it can slow down the system since the search at retrieval time has now to deal with a larger number of cases. However, if the integration of background knowledge is done intelligently, it can also help it condense the set of features that are useful in arriving at a revised representation of cases that is more reflective of their similarities. This appears more intuitive and in concordance with Dreyfus observation above. In the current paper, we are restricted to a set of features derived from the cases and from the Wikipedia articles. This is the view of the system at a given timestamp. We can extrapolate this view to a situation where the system acquires more and more cases, and as it grows in the size of the case-base, the feature set also evolves with time. The set of features can even reduce in number if we discover that all cases are about just a few underlying topics (concepts). In the context of the current paper, this implies that the cases can be meaningfully interpreted if we have access to a small number of Wikipedia articles. Since the performance of kNN-based approaches is more critically dependent on the dimensionality of the space than on the number of cases (the curse of dimensionality[6]), this progressive reduction in the dimensionality can lead to faster retrieval with fewer features.

In a general setting, the problem of determining the right set of Wikipedia articles is an optimization problem. The objective function in the supervised case corresponds to classification effectiveness averaged over the several folds created out of the training data. In an unsupervised setting, we can aim at

finding features (articles) that minimize the case-base complexity. In other words we would like to construct representations of the problem and solution components of cases such that in problems close to each other in the revised problem space, correspond to solutions that are close to each other in the revised solution space. This corresponds, for example, to a TCBR system where both problem and solution components are textual. The Wikipedia articles used to describe the problem space may be very different from the Wikipedia articles used to represent the solution space.

7 Conclusion

The effectiveness of a TCBR system is critically dependent on the representation of cases, and the measure of semantic relatedness between cases. There have been several studies into introspectively learning strategies that exploit co-occurrence patterns between words and phrases. In this paper, we present an approach to selectively exploiting background knowledge to construct richer case representations. We present empirical evidence to suggest that this approach can achieve significant effectiveness gains with very few features. We compare several article selection strategies to identify Wikipedia articles that can potentially have high impact on classification effectiveness. We also examine the effects of incorporating knowledge of relatedness between these features. We hope that the paper will stimulate further research that aim at constructing TCBR systems that are knowledge rich, easy to maintain and can be adapted to capture as much domain knowledge as is needed to suit the requirements of the specific task at hand, while compensating for the lack of cases that "cover" the problem domain adequately.

References

1. Chakraborti, S., Ambati, S., Balaraman, V., Khemani, D.: Integrating knowledge sources and acquiring vocabulary for textual CBR. In: UK-CBR Workshop, pp. 74–84 (2004)
2. Gabrowich, E., Markovith, S.: Computing semantic relatedness using Wikipedia based explicit semantic analysis. In: Proc. of Int. Joint Conference on AI, pp. 1606–1611 (2007)
3. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph, 235–244 (1990)
4. Lenz, M.: Case Retrieval Nets as a Model for Building Flexible Information Systems, PhD dissertation, Humboldt Uni. Berlin. Faculty of Mathematics and Natural Sciences (1999)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science, 391–407 (1990)
6. Mitchell, T.: Machine Learning. McGraw Hill International (1997)
7. Wiratunga, N., Lothian, R., Chakraborti, S., Koychev, I.: A propositional approach to textual case indexing. In: Proc. of European Conference on Principles and Practice of KDD, pp. 380–391 (2005)

8. Chakraborti, S., Lothian, R., Wiratunga, N., Watt, S.: Sprinkling: Supervised Latent Semantic Indexing. In: Proc. of Annual European Conference on Information Retrieval, pp. 510–514 (2006)
9. Sebastiani, F.: Machine Learning in automated text categorization. ACM Computing Surveys, 1–47 (2002)
10. Zelikovitz, S., Hirsh, H.: Using LSI for Text Classification in the Presence of Background Text. In: Proc. of International Conference on Information and Knowledge Management, pp. 113–118 (2001)
11. Scott, S., Matwin, S.: Text classification using Wordnet Hypernyms. In: Workshop on Usage of WordNet in NLP Systems, pp. 45–51 (1998)
12. Rodriguez, M., Gomez-Hidalgo, Z., Diaz-Agudo, B.: Using WordNet to Complement Training Information in Text Categorization. In: The Proc. RANLP, pp. 25–27 (1997)
13. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Machine Learning: Proceedings of the Twelfth International Conference (1995)