**On the Integration of Background Knowledge in TCBR systems**

Anil Patelia1, Sutanu Chakraborti1, Nirmalie Wiratunga2

1Department of Computer Science and Engineering, Indian Institute of Technology Madras

 Chennai: 600036. India

2School of Computing, Robert Gordon University, Aberdeen AB25 1HG, Scotland, UK

Email: pateliaaj@cse.iitm.ac.in, sutanuc@iitm.ac.in, nw@comp.rgu.ac.uk

**Abstract**. This paper explores issues in integrating background knowledge from freely available web resources like Wikipedia into Textual Case Based Reasoning. The work reported here can be viewed as an extension of an earlier paper on Explicit Semantic Analysis, which attempts at representing text in terms of concepts, where each concept has a correspondence with a Wikipedia article. We present approaches to identify Wikipedia pages that are likely to contribute positively to the effectiveness of the TCBR system in classification tasks. We also study the effect of modelling semantic relatedness between the concepts (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.

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 benefit from a principled integration of background knowledge. This paper is about some preliminary experiments we conducted towards this goal. In particular, we show how background knowledge as is readily available in resources like Wikipedia can lead to conspicuous improvements in retrieval effectiveness. We restrict the scope of this paper to classification domains.

 It may be noted here that there have been several efforts in the past outside the TCBR community that aimed at integrating background knowledge for text retrieval and classification. We have attempted a brief overview of these approaches in Section 2. One of these approaches is the Explicit Semantic Analysis (ESA) presented by Gabrilovich et al [5], which allows for easy integration of Wikipedia knowledge into instance based learners. The key idea is to treat Wikipedia articles as concepts and construct representation of documents 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, however. The first is: how do we identify the set of Wikipedia articles (concepts) relevant to a given task? The second is: 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? This paper positions ESA in the context of TCBR, and attempts to answer these questions.

 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. Section 4 describes how similarity between Wikipedia articles can be estimated, and how this knowledge can be used to obtain revised representation of cases. Section 5 presents empirical evaluation of our approaches. We conclude the report and present avenues for future work in section 6.

1. **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, one containing the word “rook” and another 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. This shows that Wikipedia knowledge can help in arriving at better models of semantic relatedness between words, as well as between documents. The key idea behind ESA[5] 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.

* 1. **A brief introduction to the ESA algorithm and its relatives**

We show in Figure 1 a schematic to illustrate how ESA can be used in Text Classification.



**Fig. 1.** Explicit Semantic Analysis

 In this approach each training document is mapped to a concept representation. Each concept corresponds to a Wikipedia article. The semantic relatedness 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 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.

 There have been other works that aimed at creating revised representations of documents based on linguistic or background knowledge. Scott et al[21] used the synonymy and hypernymy relations from WordNet [6] to revise bag-of-words representations. Rodriguez et al[22] used WordNet to enhance neural network learning algorithms for significant improvements in classification accuracy. Zelikovitz et al. [20] 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.

1. **Wikipedia Article Selection Strategies**

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. 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 only 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. This approach is often not adequate for two reasons. Firstly, not all Wikipedia pages tagged with the relevant category labels will help in discriminating between the classes. Secondly, we may still have a large number of redundant Wikipedia pages being considered. We proposed and experimented with four different Wikipedia page selection strategies with the goal of addressing these shortcomings. In this section, we briefly discuss these approaches. We have used Scrapbook Firefox (Web browser) extension to download Wikipedia articles. Scrapbook allows downloading web article with domain restriction (for example Wikipedia domain) and depth limitation. Then we deleted articles which have more common Wikipedia titles like Wikipedia:About, Main\_page, donate etc and extracted text from the remaining html articles. We have given category titles (like en.wikipedia.org/wiki/Religion) as seed to scrapbook.

**3.1. Centroid strategy.**

For each category, we compute the centroid of training documents in that category. The cosine similarity between each Wikipedia article and the centroid of each of these categories is computed. The articles are then ranked in the descending order of the maximum cosine similarity they have with any cluster centroid, and the top *k* articles are selected. The basic idea behind this approach is that web-pages prototypical of the categories should be selected. 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**

The underlying principle of this strategy is very similar to that of the centroid strategy. The significant difference is that we no longer use the centroid as a representative of a category. 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. 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 category centroid. A limitation of this could be that Wikipedia pages that are extremely similar to each other can get selected, leading to redundancy.

**3.3. k-Nearest Neighbour with Discrimination**

This strategy is very similar to the kNN approach, except that we ensure that each category gets its share of representative articles. Thus we select the top-k Wikipedia articles from each category that have higher cosine similarities with the three closest neighbours in training documents of that category. This overcomes the second limitation of the Centroid strategy in that it guarantees that no category suffers from starvation.

**3.4**. **Probability Ratio strategy**

This strategy evaluates the relative importance of a Wikipedia article to its category in terms of a ratio of posterior probabilities. This approach is founded on probability estimates computed from the training corpus as in Naïve Bayes style classifiers using add-1 smoothing. Given a category ci drawn from a set of *n* categories, such that ci є {c1, c2, …, cn}, the posterior probability P(ci|wk) of ci given a Wikipedia articles wk is given by P(ci|wk). We calculate the ratio P(cj|wk) /  for each Wikipedia article wk and assign these articles to the categories based on these scores.

 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 want to miss out on those words that are already good in discriminating between classes. We name this approach Augmented ESA Representation.

**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 subsection, we discuss how we incorporate the knowledge of similarity between Wikipedia articles into the revised representation of terms and documents.



**Fig. 2.** Case Retrieval Network

 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 relevancies 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 2. 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 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 2. If the aggregation function at each node is a simple summation, the resulting representation of D out to be {1.9, 1.9, 1.1, 0, 0}. This new representation can be seen as a result of a matrix operation. Let Ri and Rn 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 Rn = RiS.

 The similarities between Wikipedia articles are estimated using Latent Semantic Analysis. This is in line with an earlier work where LSA was used for introspective knowledge acquisition in CRNs[3]. The basic motivation in using LSA for estimating term pair similarities is its ability to model higher order co-occurrences between terms. LSA takes as input a term document matrix and aims at constructing a reduced dimensional vector space spanned by a set of orthogonal concepts, in terms of which both documents and words can be represented. The underlying philosophy is Dual Mode Factor Analysis, with the basic hypothesis that a concept based representation helps in weeding out “noise” due to word choice variability. An intuitive introduction to the Singular Value Decomposition which constitutes the essential apparatus of LSA is presented in [9]. For the purpose of the current discussion, it is sufficient to note that once words are represented in the lower dimensional concept space the computation of similarities between words amounts to taking a cosine similarity between the word vectors in the concept space.

 A further refinement to the idea of using LSA for similarity knowledge acquisition is the idea of sprinkling[16], 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 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 [16].

**5 Evaluations**

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 [12] 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 corpuses were formed from the news-group: 1. HARDWARE group from two hardware categories, one on MAC and the other on PC. 2. RELPOL, from two groups, one concerning religion, the other politics in the middle-east. 3 SCIENCE from four science related groups 4. REC from four recreation related groups. Thus HARDWARE and RELPOL are two class problems, and SCIENCE and REC are four-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 *k* NN classifier is used with *k* = 3.

**5.2. Identifying relevant Wikipedia Articles**

In this subsection, we show top ten Wikipedia articles selected under Religion and Politics categories in the RELPOL dataset using various article selection strategies described in section 3. Table 1 shows the title of top ten Wikipedia articles under the Religion category. All these were extracted from Wikipedia pages labelled with the Religion category. Similar data for Politics is shown in Table 2.

**Table 1.** Top 10 Wikipedia articles by four different selection strategies for Religion category

|  |  |
| --- | --- |
| doc Rank | Wikipedia document Selection strategy |
| Centroid | Knn | knnDiscri | probRatio |
| 1 | Jesus | Salvation | Apophatic\_theology | holy\_trinity |
| 2 | jesus\_christ | Trinitarian | Trinitarian | Trinitt |
| 3 | jesus\_of\_nazareth | Theoria | Binitarianism | Sabbath |
| 4 | Gnosticism | Sabbath | divine\_grace | Deism |
| 5 | Gnostics | Gnosticism | Salvation | Deist |
| 6 | holy\_trinity | Gnostics | justification\_(theology) | Deists |
| 7 | Trinity | Apophatic\_theology | Fideism | Gnosticism |
| 8 | Predestination | Jesus | Irresistible\_grace | Gnostics |
| 9 | Deism | jesus\_christ | imputed\_righteousness | Salvation. |
| 10 | Deist | jesus\_of\_nazareth | existence\_of\_god | Trinitarian |

**Table 2.** Top 10 Wikipedia articles by four different selection strategies for Politics category

|  |  |
| --- | --- |
| doc Rank | Wikipedia document Selection strategy |
| centroid | Knn | knnDiscri | probRatio |
| 1 | first\_world\_war | Hungary | korean\_war | first\_world\_war |
| 2 | world\_war\_i | adolf\_hitler | league\_of\_nations | world\_war\_i |
| 3 | adolf\_hitler | first\_world\_war | first\_world\_war | iraq\_war |
| 4 | nazi\_germany | world\_war\_i | world\_war\_i | civil\_war |
| 5 | third\_reich | Russia | Russia | korean\_war |
| 6 | civil\_war | league\_of\_nations | Hungary | Terrorism |
| 7 | joseph\_stalin | winston\_churchill | Afghanistan | Vietnam\_war |
| 8 | iraq\_war | Afghanistan | Turkey | war\_on\_terrorism |
| 9 | vietnam\_war | vietnam\_war | India | united\_states |
| 10 | Hungary | korean\_war | Haiti | India |

As shown in Tables 1 and 2, the Wikipedia articles are related to their categories religion and politics respectively. But we have observed that Wikipedia articles like personal\_computer.txt 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.

 The Baseline algorithm used for our comparisons is one 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. Table 3 compares the accuracies of ESA against a naïve bag-of-words Vector Space approach and the Baseline on 4 sub category from the Twenty Newsgroup. Table 4 reports the accuracies obtained when Augmented ESA representation (refer Section 3) 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 ad hoc selection procedure.



**Fig. 3.** Classification accuracy in percentage as a function of background knowledge size (no. of wiki documents) for HARDWARE Dataset



**Fig. 4.** Classification accuracy in percentage as a function of background knowledge size (no. of wiki documents) for RELPOL Dataset

 The general trend observed in these graphs is that classification accuracy increases as a function of the number of Wikipedia pages. As shown in figure 3 and figure4, for both RELPOL and HARDWARE, the increase is steeper than with random selection of Web pages. This shows that we can attain conspicuous improvements using fewer pages, if we adopt a principled approach to selection of Wikipedia articles.

**5.3 Modeling similarity between Wikipedia pages**

 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 5 shows the classification accuracy using the revised case representation which incorporates knowledge of similarities between concepts. Table 6 shows the results when Augmented ESA representation is used, along with knowledge of similarities between concepts.

**Table 3.** Comparison of performance of ESA, Vector Space Model (VSM) and Baseline on 20 Newsgroup

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Wiki article selection strategy | VSM | Baseline |
| Centroid | knn | knnDiscri | probRatio |
|  HARDWARE | 77.72 | 76.51 | 74.60 | 80.28 | 59.51 | 65.77 |
|  RELPOL | 92.94 | 92.98 | 92.43 | 92.54 | 70.51 | 80.88 |
|  SCIENCE | 81.01 | 76.76 | 78.34 | 77.98 | 54.89 | 60.22 |
|  RECREATION | 83.02 | 76.76 | 79.72 | 77.26 | 62.79 | 66.54 |

**Table 4.** Pperformance of Augmented ESA on 20 Newsgroup

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Wiki article selection strategy |  VSM |  |
| Centroid | Knn | knnDiscri | probRatio | Baseline |
|  HARDWARE | 74.75 | 76.70 | 76.51 | 75.84 | 59.51 | 65.77 |
|  RELPOL | 93.13 | 93.09 | 93.04 | 93.09 | 70.51 | 80.88 |
|  SCIENCE | 77.76 | 78.16 | 76.44 | 77.63 | 54.89 | 60.22 |
|  RECREATION | 82.68 | 77.45 | 77.31 | 79.23 | 62.79 | 66.54 |

**Table 5.** Performance of ESA with knowledge of concept similarities on 20 Newsgroup

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Wiki article selection strategy | VSM | Baseline |
| Centroid | Knn | knnDiscri | probRatio |  |
|  HARDWARE | 75.37 | 75.37 | 72.38 | 78.12 | 59.51 | 65.77 |
|  RELPOL | 93.08 | 92.64 | 91.04 | 92.49 | 70.51 | 80.88 |
|  SCIENCE | 79.69 | 77.91 | 76.13 | 75.00 | 54.89 | 60.22 |
|  RECREATION | 80.05 | 72.89 | 76.43 | 75.66 | 62.79 | 66.54 |

**Table 6.** Performance of Augmented ESA with knowledge of concept similarities on 20 Newsgroup

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Wiki article selection strategy | VSM | Baseline |
| Centroid | Knn | knnDiscri | probRatio |
|  HARDWARE | 76.57 | 74.23 | 74.23 | 75.77 | 59.51 | 65.77 |
|  RELPOL | 94.49 | 94.31 | 93.66 | 94.23 | 70.51 | 80.88 |
|  SCIENCE | 82.49 | 80.83 | 79.95 | 79.78 | 54.89 | 60.22 |
|  RECREATION | 80.87 | 78.56 | 77.82 | 79.89 | 62.79 | 66.54 |

**5.4 Summary of Observations**

We observe that after integration of background knowledge, effectiveness of text classification improves conspicuously, and the differences vary from 5% to 20%. As shown in Table 3, Baseline algorithm performs better than naïve Vector Space model for each category and differences vary from 3% to 10%. Paired t-test was carried out between accuracies reported by each pair of methods over the 15 train test pairs; Baseline algorithm outperform the vector space model at p = 0.05. And each Wikipedia article selection strategy performs better than Baseline for each category outperform the Baseline at p = 0.05 in paired t-test. 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 3. Comparing the Wikipedia article selection strategies, centroid is better than remaining three. As shown in Table 4, Augmented ESA performs better than ESA for RELPOL category. Similarity modelling between Wikipedia articles does not improve the result for ESA presented in Table 5. In particular, the highest accuracy is 94.49% when centroid strategy is used for augmented ESA representation with similarity modelling. Our analysis shows that Augmentation ESA with similarity modelling performs better than remaining case representation using background knowledge. Augmented ESA with similarity modelling performs better than other case representation for SCIENCE dataset for each Wikipedia articles selection strategies. For HARDWARE dataset, performance decreases as we try to model similarity between Wikipedia articles. This may be due to there are many common wiki articles in Wikipedia articles collection for both category. There is no significant difference in all text representation with every Wikipedia page selection strategy except few cases at p = 0.05 in paired t-test. For example, In RELPOL dataset augmented ESA with similarity modelling performs better than ESA with similarity modelling at p = 0.05 in paired t-test. We observe that integration of background knowledge results in significant improvement in TCBR.

. We have tried different number of Wikipedia articles for ESA representation, and as we increased the number of articles, the accuracy increases monotonically with some exceptions as shown in figure 3 and 4. As shown in figure 3, ESA representation with 20 wikipedia articles in page selection strategy performs minimum 8% better than vector space model which performs at 70.51% in each document selection strategy.

 Support Vector Machines over a linear kernel gives the accuracy 78.82% for hardware category and 91.86% for RELPOL in vector space model. Explicit Semantic Analysis with probability-ratio article selection strategy gives 80.28% accuracy which is better than SVM on its own. In RELPOL, most article selection strategies outperform SVM. The results presented indicate that SVM over ESA with Augmented representation outperforms SVM over VSM and baseline.

**6 Conclusion and Future Work**

The empirical studies presented in this paper demonstrate that a principled integration of background knowledge into TCBR tasks can lead to significant improvements in the effectiveness. While resources like Wikipedia are useful, it is important to identify that small subset of pages that are expected to actually contribute positively to the task in question. We have presented and compared approaches for selection of Wikipedia pages relevant to text classification. The results show a significant improvement over a baseline approach that randomly selects pages from relevant categories. We have also conducted a preliminary study of the effects of modelling semantic relatedness between concepts (as represented by Wikipedia articles) and shown that this knowledge can lead to further improvements. We need to conduct further investigations to identify similarity relationships that do not contribute to improving discrimination between classes. This may help arrest the decline in performance in datasets like hardware, where the articles belonging to the two distinct categories have a lot of overlap in word usage. Unlike earlier work, we have experimented with a mix of features that include the words as well as concepts obtained from Wikipedia for text representation. This approach outperforms an approach that only represents text using Wikipedia concepts. Though the results have been presented in the context of classification tasks, extensions can be made to non-classification tasks like retrieval or question answering as well.

 As part of future work we plan to take a fresh look at the process of background knowledge from a cognitive science perspective. It would be interesting to study how humans integrate statistical, linguistic and background knowledge as they make sense of text. The integration appears to be on an on-demand basis, with each piece of knowledge called in to resolve conflicts generated by the other sources. Simple tasks may not involve a lot of switching between distinct sources; while more involved tasks may demand a more complex interaction. Hyperlink structure in Wikipedia articles can be used as an additional knowledge source to improve textual case representation.

**References**

## David W.  A., and Héctor M. A.: Introduction: Interactive Case Based Reasoning. Applied Intelligence, 14 1 2001, pp. 1573-7497

1. Chakraborti, S., Beresi, U. C., Wiratunga, N., Massie, S., Lothian, R., and Khemani, D.: Visualizing and Evaluating Complexity of Textual Case Bases. Proc of 9th European conference on advances in CBR , 2008 , pp. 104-119
2. Chakraborti, S., Ambati, S., Balaraman, V., and Khemani: Integrating knowledge sources and acquiring vocabulary for textual CBR, Proc of 8th UK-CBR workshop, 2004, pp. 74-84
3. Diaz-Agudo, B., and González-Calero, P. A.: Formal Concept analysis as a support technique for CBR, Knowledge Based Syst, 14(3-4), 2001, pp. 163-171
4. Gabrowich, E., and Markovith, S.: Computing semantic relatedness using Wikipedia based explicit semantic analysis, Proc of 20th Int joint conference on AI, 2007, pp. 1606- 1611
5. Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., and Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph, 3 4 1990, pp. 235-244.
6. http://en.wikipedia.org
7. 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)
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R.: Indexing by latent semantic analysis. Journal of the American soceity for Information Science 41(6) (1990), pp. 391-407
9. Kontostathis, A., and Pottenger, W. M.: Detecting patterns in LSI term-term matrix. Proc of IEEE ICDM02 workshop, pp. 243-248
10. Raghunandan, M. A., Chakraborti, S., and Khemani, D.: Robust Measures of Complexity in TCBR, Proc of 8th International Conference on Case-based reasoning (2009)
11. Mitchell, T: Machine Learning. Mc Graw Hill International,1997
12. Vapnik, V. N.: Statistical Learning theory. Wiley , 1998
13. Wiratunga, N., Lothian, R., Chakraborti, S., and Koychev, I.: A propositional approach to textual case indexing, Proc of 9th European Conference on Principles and Practice of KDD, 2005, pp. 380-391
14. Friedman, J., Hastie, T., and Tibshirani, R.: Additive Logistic Regression : A statistical view of Boosting, Annals of Statistics, 28(2), pp. 337-407
15. Chakraborti, S., Lothian, R., Wiratunga, N., and Watt, S.: Sprinkling: Supervised Latent Semantic Indexing. In the proc of Annual European Conference on Information Retrieval, 2006, pp. 510-514
16. Joachims, T.: Making Large-Scale SVM Learning Practical. Advances in Kernel Methods: Support Vector Learning, 1999
17. Sebastini, F.: Machine Learning in automated text categorization. ACM Computing Surveys, 34(1), 2002, pp. 1-47
18. Bruce, W. P., Ray, B., and Robert, C. H.: Concept Learning and Heuristic Classification in Weak-Theory Domains. Readings in Knowledge Acquisition and learning: automating the construction and improvement of exp. Systems., 1993, pp. 741-758
19. Zelikovitz S., and Hirsh, H.: Using LSI for Text Classification in the Presence of Background Text. International Conference on Information and Knowledge Management, 2001,pp. 113- 118

21. Sam, S., and Stan, M.: Text classification using Wordnet Hypernyms, Workshop on usage of

 WordNet in NLP systems, 1998, pp. 45-51

22. Manuel de Buenaga Rodriguez, Zas 6 Marfa G6mez-Hidalgo and Ben Dfaz-Agudo. Using WordNet to Complement Training Information in Text Categorization. In Proc. RANLP-97, Stanford, March 25-27, 1997