Semantic Characterisation: Knowledge Discovery for Training Set

Tan Ping Ping, Narayanan Kulathuramaiyer, and Azlina Ahmadi Julaihi

***Abstract*—This paper has proposed the use Latent Semantic Indexing (LSI) to extract semantic information to make the best use of the existing knowledge contained in training sets: Semantic Characterisation (SemC). SemC uses LSI to capture the implicit semantic structure in documents by directly applying category labels imposed by experts to make semantic structure explicit. The training set filtered by SemC is tested on a supervised automated text categorisation system using Support Vector Machine as classifier. Category by category analysis has shown the ability to bring out the semantic characteristics of the datasets. Even with a reduced training set, SemC is able to overcome the generalisation problem due to its ability to reduce noise within individual categories. Our empirical results also demonstrated that SemC managed to improve categorisation results of heavily overlapping categories. Empirical results also showed that SemC is applicable to a various supervised classifiers.**

***Index Terms*—Automated text categorisation, dataset, latent semantic indexing.**

I. Introduction

Even though automated text categorisation (ATC) attempts to mimic the categorisation model of human experts, current supervised ATC systems tend to merely exploit the information captured from a set of pre-determined documents class label assignments. During the process of human assignment of class labels, a great deal of implicit knowledge is employed. This knowledge is however not made explicit and systematically captured during the process of manually classifying documents. Reasons why a document is assigned to a particular class, if captured effectively, could provide valuable information for knowledge intensive tasks such as text categorisation.

When a document is determined to belong explicitly to exactly one category, (which is typical for most classified datasets) single-class category labels are assigned despite of document content overlaps with other categories. The existence of content overlaps across classes tends to complicate learning process [1]. Efforts that have been taken to overcome this problem [2] mainly concentrate on very specific domains or involves the addition of unlabelled training sets, thus, causing overfitting.

This paper explores the discovery of intrinsic patterns and characteristics of datasets in an effort to determine the characteristics of datasets that influence the performance of classifiers. Latent Semantic Indexing (LSI) [3] has been applied as a means of extraction of latent document dependency structures [4]. The relationships between extracted information about datasets are explored through an analysis of text categorisation results.

II. Semantic Characterisation: Algorithm

```
Semantic Characterisation algorithm:
For training set, $T_r$ with predefined categories, $C$
Where category labels, $C_i$; $i$ = number of categories in $T_r$
Let $T_c$ = set of positive examples for $C_\mathrm{i}$
Let $\hat{X}$ =Reduced singular value decomposition of the term x document matrix for $T_r$,
For category $C_i$,
Using $C_i$ as query (query terms selection according to the validity),
LSI retrieval is performed on (by ignoring the predefined labels)
Let $T_l$ = set selected by LSI
Let $T_s = T_l\ T_c$ which represents the positive examples in $T_r$ and LSI
Let $T'$ = $T_s$ for all $C$
Train supervised classifier $h$ using Na ïve Bayes / SVM on $T'$
```

This work differs from other works that employ LSI, whereby LSI is not used as a feature extraction method or by manipulating LSI's vector spaces like existing supervised LSI methods [5] – [8], SemC, on the other hand, makes use of the existing categorical information and LSI's retrieval method; query technique and manipulation of the retrieval results of LSI to re-model the training sets. Thus, our approach explicates the meaning of training sets and queries applied to become directly interpretable by users while uncovering valuable category knowledge used by experts when performing document classification.

The knowledge contained in the training sets can then be manipulated through the selection of query terms. Hence, this eliminates the need to perform singular value decomposition locally for each separate category. SemC as such, does not require additional knowledge to be elicited from experts, as it is able to make use of latent document-term distribution patterns as contained the training set. The application of SemC results in a significantly reduced training set derived from the intrinsic text content characteristics of the training set.

SemC has been tested on a probabilistic classifier: multinomial Na ïve Bayes (MNB) and has shown promising results [9]. This paper then reports the findings of experimentations in applying SemC in conjunction with the SVM classifier.

Manuscript received September 22, 2012; revised November 13, 2012. This work was supported in part by Universiti Malaysia Sarawak.

The authors are with the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Malaysia (e-mail: pptan@fit.unimas.my, nara@fit.unimas.my, ajazlina@fit.unimas.my).

DOI: 10.7763/IJIMT.2013.V4.357

III. Methodology

The experiments were conducted comparing SemC's filtered training set contrasting it with the full training set used as a baseline. For SemC filtering, the Reuters-21578 [10] top ten categories training set were prepared for LSI retrieval using the General Text Parser (GTP) system [11] with category labels used as queries. Then the retrieved documents are placed for SemC filtering. SemC's filtered category by category training sets were combined to form a complete training set with positive and negative examples. The reduced training set was fed to the text processing system [12] as training set. For the baseline the full training set was used by the ATC system as training set. The testing set used for both experiments were the same, using original documents from the dataset without perfoming SemC filtering.

The Waikato Environment for Knowledge Analysis (WEKA) [13] was used for Machine Learning (ML). WEKA's sequential minimal optimization model of SVM was used for the experiments. The categorisation results using SemC's filtered training set was then compared directly with the results from the full training set.

IV. Results and Analysis

The effectiveness of LSI as a retrieval method is shown in Table I. The overall categorisation result is shown in Fig. 1. SemC's degree of reduction was affected by the intrinsic characteristics of the dataset for each of the categories. Although SemC's reduced training set represents a subset of features from the original training set, the training feature spaces produced for both sets have differing terms occurrences due to differences in the number of training set documents. These training sets then produced different margins of separations for positive and negative examples.

Incorporating the latent semantic structure and class information of the dataset, SemC was able to isolate meaningful documents and has shown to eliminate noise in the training set. This shows that by employing the latent semantic structure, as a basis for training set reduction, the overall categorisation performance increases regardless the number of training set used. This has resulted in a compact representation of the categories without compromising on categorisation quality.

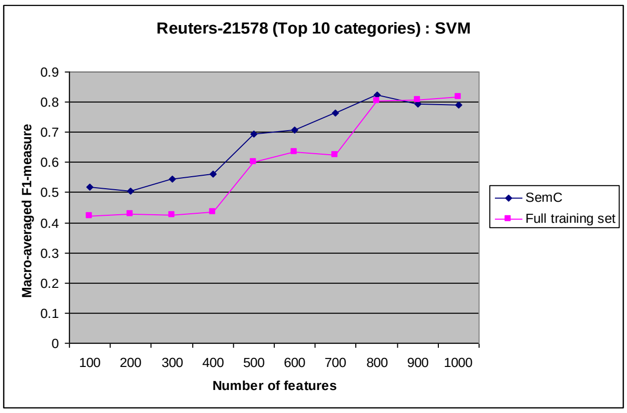
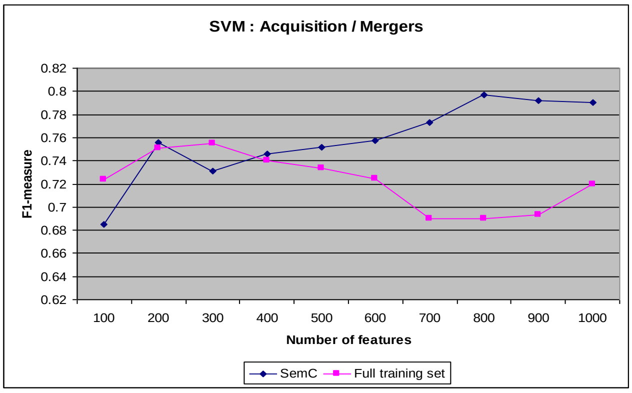


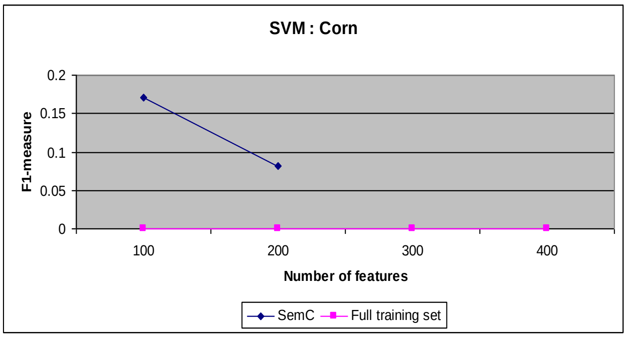
Fig. 1. Overall categorisation results.

TABLE I: SemC Filtering Results on Reuters-21578 Top Ten Most Common Categories.

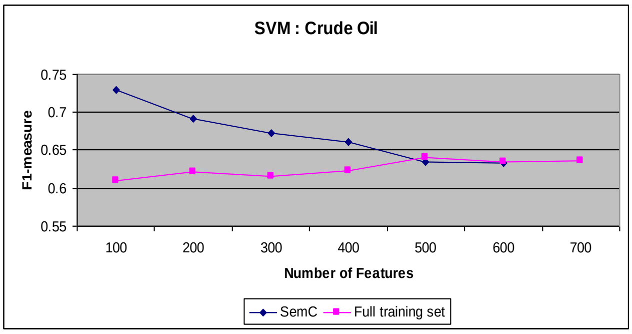
Category	No. of Training Documents		Number of Documents Retrieved by LSI	Reduction (%)
	SemC	Original		
Acquisition / merger	1248	1650	1434	24.4
Corn	109	181	390	39.8
Crude oil	346	387	753	10.6
Earnings and earning forecast	503	2862	629	82.4
Grain	153	429	445	64.3
Interest rate	229	345	624	33.6
Money foreign exchange	312	535	852	41.7
Shipping	164	192	529	14.6
Trade	336	367	781	8.5
Wheat	188	212	603	11.3
Total	3588	719		49.9



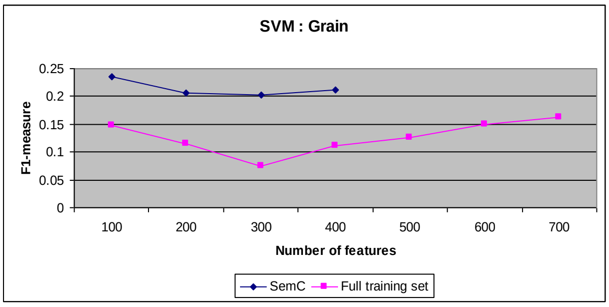
(a) Category 'Acquisition / Mergers'



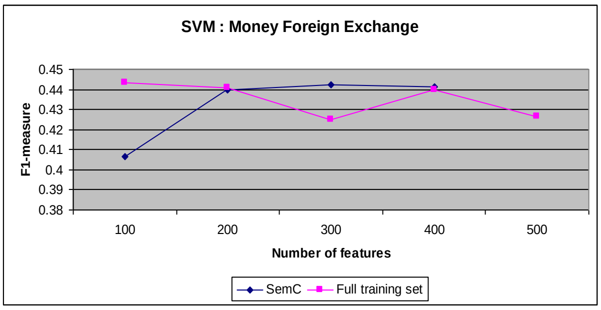
(b) Category 'Corn'



(c) Category 'Crude Oil'



(d) Category 'Grain'



(f) Category 'Money foreign exchange'

Fig. 2. The patterns of results observed where SemC were able to capture the intrinsic characteristics of the training set.

By eliminating ambiguous training documents in SemC, SVM was able to form the optimal feature set in supporting the categorisation of new testing set documents. SVM has demonstrated its strength: high dimensional feature spaces, few irrelevant features (dense concept vector), and sparse instance vectors [14].

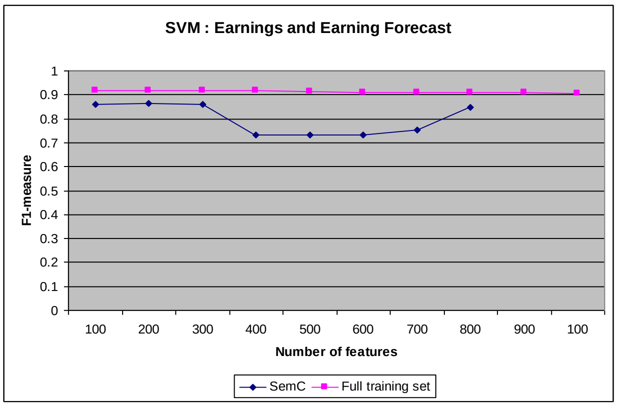


Fig. 3. Category 'earnings and earning forecast'.

A further observation of the category by category empirical results has shown that, by employing a smaller set of features, SemC was able to perform significantly better for 9 out of 10 categories compared to using full training set (Fig. 2). When the reduction degree was very high, SemC performed marginally worse for a category (Figure 3). Our manual analysis of the document showed that for Miscallaneous categories, SemC's high level reduction causes SemC to filter out distinct examples in the training set. By having too high of a reduction also, negative examples tend to overshadow positive examples making positive examples less dominant.

V. Conclusion

Hence, our proposed method: Semantic Characterisation (SemC) has shown to be able to extract semantic information to make better use of the semantic information employed by human which reflects an experts' mental model. In other words the application of implicit information about document assignment to category can enhance performance of classifiers. The benefit of this approach to supervised ATC systems is in the resulting reduction in training set documents. SemC addresses problem generally faced by supervised ATC like class relationships, closely related documents (synonyms words) and acquisition of semantic meanings in datasets.

Acknowledgment

The authors thank the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak.

References

[1] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Exploration*, vol. 6, no. 1, pp. 20-9, 2004.

[2] X. Lu, B. Zheng, A. Velivelli, and C. X. Zhai, "Enhancing text categorization with semantic-enriched representation and training data augmentation," *Journal of American Medical Informatics Association*, vol. 13, no. 5, pp. 526-35, Sep-Oct. 2006.

[3] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. (1990) Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6), pp. 391-407. [Online]. Available: http://lsi.research.telcordia.com/lsi/papers/JASIS90.pdf

[4] D. Bassu and C. Behrens, "Distributed LSI: Scalable concept-based information retrieval with high semantic resolution," in *Proceedings of the 3rd SIAM International Conference on Data Mining (Text Mining Workshop)*, San Francisco, CA, May 3, 2003.

[5] S. Zelikovitz and F. Marquez, "Evaluation of background knowledge for latent semantic indexing classification," *American Association for Artificial Intelligence*, 2005

[6] S. Chakraborti, R. Mukras, R. Lothian, N. Wiratunga, and D. W. S. Harper, "Supervised latent semantic indexing using adaptive sprinkling," *International Joint Conference on Artificial Intelligence Hyderabad*, India, pp. 1582-1587, January 2007.

[7] Y. Qing and L. F. Min, "Support vector machine for customized email filtering based on improving latent semantic indexing," *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics Guangzhou*, China, pp. 18-21, August 2005.

[8] J. T. Sun, Z. Chen, H. J. Zeng, Y. C. Lu, C. Y. Shi, and W. Y. Ma, "Supervised latent semantic indexing for document categorization," *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)*, 2004.

[9] P. P. Tan and N. Kulathuramaiyer, "Latent semantic indexing for training set reduction conference paper categorization," *Proceedings of the 4th International Conference on Information Technology* in Asia Malaysia, 2005.

[10] D. D. Lewis, Reuters-21578 Distribution 1.0. 26 September 1997

[11] S. W. Ling, J. Giles, and M. W. Berry, *General Text Parser (GTP) Software.* 2003.

[12] C. H. Bong, K. Narayanan, and P. P. Tan, "Text mining workbench," *Poster Paper in IPTA Research and Development Exhibition and Conference* Malaysia, 2003.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[14] T. Joachims, "Text categorization with support vector machine: learning with many relevant features," *Proceedings of the European Conference on Machine Learning Springer*, 1998.