A Dimensionality Reduction Approach for Semantic Document Classification

Oskar Ahlgren, Pekka Malo, Ankur Sinha, Pekka Korhonen & Jyrki Wallenius

Aalto University School of Economics
P.O. Box 21210, FI-00076 AALTO, FINLAND

Abstract. The curse of dimensionality is a well-recognized problem in the field of document filtering. In particular, this concerns methods where vector space models are utilized to describe the document-concept space. When performing content classification across a variety of topics, the number of different concepts (dimensions) rapidly explodes and as a result many techniques are rendered inapplicable. Furthermore the extent of information represented by each of the concepts may vary significantly. In this paper, we present a dimensionality reduction approach which approximates the user's preferences in the form of value function and leads to a quick and efficient filtering procedure. The proposed system requires the user to provide preference information in the form of a training set in order to generate a search rule. Each document in the training set is profiled into a vector of concepts. The document profiling is accomplished by utilizing Wikipedia-articles to define the semantic information contained in words which allows them to be perceived as concepts. Once the set of concepts contained in the training set is known, a modified Wilks' lambda approach is used for dimensionality reduction by ensuring minimal loss of semantic information.

1 Introduction

Most information retrieval systems are based on free language searching, where the user can compose any ad hoc query by presenting a list of keywords or a short phrase to describe a topic. The popularity of phrase-based methods can largely be explained by their convenience for the users. However, the ease of usage comes with a few drawbacks as well. While exploring a new topic or searching within an expert domain with specialized terminology it can be surprisingly hard to find the right words for getting the relevant content. To cope with the ambiguity of the vocabulary, concept-based document classification techniques have been proposed, as concepts by definition cannot be ambiguous. However, the use of concepts instead of keywords is only part of the solution. If the filtering methods rely on vector-space models of documents and concepts, a dimension reduction technique comes in handy. Instead of training the classifiers using the entire concept-base, the learning of filtering models is improved by restricting the space to those concepts that are most relevant for the given task.

In this paper, we introduce, Wilks-VF, a light-weight concept selection method inspired by Wilks' lambda to reduce the curse of dimensionality. In Wilks-VF the

document classification task is carried out in the following three stages: 1) Once the user has supplied a training sample of relevant and irrelevant documents, a semantic profiler is applied to build a document-concept space representation. The semantic knowledge is drawn from Wikipedia, which provides the semantic relatedness information. 2) Next, the Wilks' lambda based dimension reduction method is used to select concepts that provide the best separation between relevant and irrelevant documents. 3) Finally, the value function framework proposed by Malo et al. [1] is employed to learn a classification rule for the given topic.

The main contribution of the Wilks-VF, as compared to the existing literature, is a light-weight concept selection method, where a clustering based Wilks' lambda approach is used to equip the methodology for on-line usability. Evaluation of the framework's classification performance is carried out using the Reuters TREC-11 corpus. The result is then benchmarked with other well-known feature selection methods. As primary performance measures we use F-Score, precision and recall. The obtained results are promising, but the work is still preliminary and further evaluation with other corpora needs to be carried out.

2 Related Work

During the last decade, the role of document content descriptors (words/phrases vs. categories/concepts) in the performance of information retrieval systems has piqued considerable interest [2]. Consequently, a number of studies have examined the benefits of using concept hierarchies or controlled vocabularies derived from ontologies and folksonomies [3] [4] [6]. In particular, the use of Wikipedia as a source of semantic knowledge has turned out to be an increasingly popular choice, see e.g. [7] [8] [9] [10] [11] [12]. The use of value function in preference modeling is well founded in the fields of operations research and management science [13] [14] [15] [16] [17]. There the purpose is to develop interactive methods for helping the users to find preferred solutions for complex decision-making problems with several competing objectives. The existence of a value function which imitates a decision maker's choices makes the two problems very similar. The essential difference between decision making problems and document classification is the high-dimensionality of information classification problems, which leads to concerns about the ability of value function based methods to deal with large number of attributes.

To alleviate the curse of the dimensionality problem which is often encountered in classification tasks, such as document filtering, a number of feature selection techniques have been proposed [18] [19] [20] [21] [22] [23] [24]. For an extensive overview of the various methods, see e.g. Fodor [25]. However, most of these techniques are designed for general purposes, whereas the approach suggested in this paper is mainly suited for concept-selection task.

3 Wilks-VF Framework

This section describes the steps of the procedure and the implementation of the framework. First, we describe the Wikipedia based document indexing procedure, where every document is transformed into a stream of concepts. Next, we present the dimensionality reduction approach utilizing Wilks' lambda, and finally we describe an efficient linear optimization method for learning the value function based document classifier.

3.1 Document profiling

The document profiling approach used in this paper is similar to the technique adopted by Malo et al.[1], where each document is profiled into a collection of concepts. To illustrate this idea, consider the example in Fig. 1. The text in the figure on the left-hand-side is transformed into a vector of concepts by the profiler. The profiler is implemented as a two-stage classifier, where disambiguation and link recognition are accomplished jointly to detect Wikipedia-concepts in the documents and each concept corresponds to a Wikipedia article. On the right-hand-side (under concept space), a small network is shown, corresponding to the central concepts found in the document. In addition to the concepts directly present in the document, the network displays also some other concepts that are specified in the Wikipedia link structure. As discussed by [7] [8] [9] [10] [11] [12] the link structure can be used for mining semantic relatedness information, which is useful for constructing concept-based classification models.

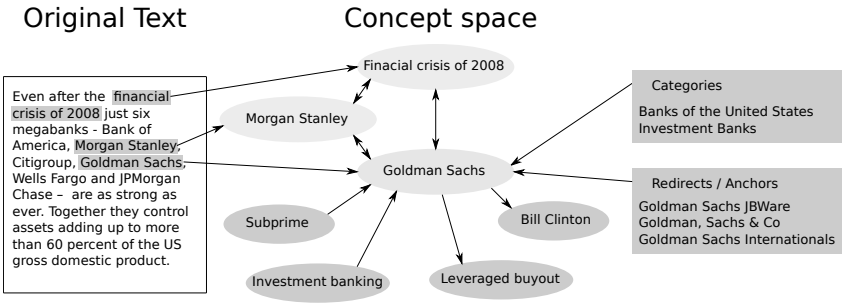

Fig. 1. Wikipedia link structure example

In this paper, we used the concept-relatedness measure described by Malo et al.[1], which in turn is inspired by the Normalized Google Distance approach proposed by Cilibrasi and Vitanyi [26]. In the following definition we introduce the concept relatedness measure and thereafter discuss its usage in Sect. 3.3.

Definition 1. *Concept relatedness: Let C denote concept-space and c_1 and c_2 be an arbitrary pair of Wikipedia-concepts. If $C_1, C_2 \subset C$ denote the sets of all articles that link to c_1 and c_2, respectively, the concept-relatedness measure is given by the mapping c-rel: $C \times C \rightarrow [0, 1]$,*

$$c\text{-}rel(c_1, c_2) = e^{-ND(c_1,c_2)},$$

where the ND measure is $ND(c_1, c_2) = \frac{\log(\max|C_1|.|C_2|)-\log(|C_1 \cap C_2|}{\log(|C|)-\log(\min(|C_1|,|C_2|)}$.

3.2 Dimension Reduction with Wilks' lambda

Wilks' lambda is used to identify the concepts which best separate the relevant documents from the irrelevant ones. Once the documents have been profiled into a matrix where each row represents a document and each column represents a concept, Wilks' lambda tests whether there are differences between the means of the two identified groups of subjects (relevant and irrelevant documents) on a number of dependent variables (concepts). A large difference in the means indicates that the chosen concept can be used to distinguish a relevant document from an irrelevant one [27].

Wilks' lambda statistic. Let $X \in \mathbb{R}^{N \times |\hat{C}|}$ denote the document-concept matrix, where N is the number of documents in the training set and $|\hat{C}|$ is the number of different concepts found in the documents. The matrix can be decomposed into two parts according to the relevance of the documents

$$X = \begin{bmatrix} X_R \\ X_{IR} \end{bmatrix},$$

where X_R and X_{IR} are the collections of profiles corresponding to the relevant documents and the irrelevant ones respectively. The dimensionality reduction procedure is based on the assumption that the profile means, $\bar{x}_R$ and $\bar{x}_{IR}$ in the two document groups are different. If $\bar{x}_R = \bar{x}_{IR}$, none of the concepts are able to differentiate between relevant and irrelevant concepts. The hypothesis $H_0 : \bar{x}_R = \bar{x}_{IR}$ can be tested by the principle of maximum likelihood, using a Wilks' lambda statistic, where T denotes the total centered cross product and W denotes the within groups cross product

$$\Lambda = \frac{|W|}{|T|} = \frac{|X_{\mathrm{R}}^T H X_{\mathrm{R}} + X_{\mathrm{IR}}^T H X_{\mathrm{IR}}|}{|X^T H X|}.$$

In the above equation, Λ follows the F-distribution and can be tested with the approximation developed by Rao [5].

Additional information. In order to choose the concepts that provide the best separation between the two groups, we employ Wilks' lambda to evaluate the information content of the concepts. Let

$$T = \begin{bmatrix} T_{11}\ T_{12} \\ T_{21}\ T_{22} \end{bmatrix} \text{ and } W = \begin{bmatrix} W_{11}\ W_{12} \\ W_{21}\ W_{22} \end{bmatrix}$$

be the block-representation of the total and between groups matrices, where $T_{11} = T_{11(q,q)}$ and $W_{11} = W_{11(q,q)}$ refer to those variables, q, that are included

into the model. For simplicity, we can assume that the variables in the model have indices $1, 2, \ldots, q$. For this purpose, we introduce the decomposition of Wilks' lambda into two parts: the information content of the selected concepts Λ_1 and the information content of the remaining concepts $\Lambda_{2,1}$:

$$\Lambda = \Lambda_1 \Lambda_{2,1} = \frac{|W_{11}|}{|T_{11}|} \frac{|W_{22} - W_{21} W_{11}^{-1} W_{12}|}{|T_{22} - T_{21} T_{11}^{-1} T_{12}|}.$$

The parts of the decomposition can be interpreted as follows:

1. if $\Lambda_1 \approx 1$, then variables $i = 1, 2, \ldots, q$ are not able to separate the groups
2. if $\Lambda_1 << 1$, then variables $i = 1, 2, \ldots, q$ separate the groups very well
3. if $\Lambda_{2,1} \approx 1$, then variables $i = q+1, q+2, \ldots, p$ are not able to provide additional information
4. if $\Lambda_{2,1} << 1$, then variables $i = q+1, q+2, \ldots, p$ contain at least some additional information

Selection heuristic. Motivated by the Wilks' lambda statistic, we now introduce the following heuristic for concept selection:

1. **Initiation**: Let $M = \{1, 2, \ldots, |C|\}$ denote the index set of all concepts and let $N = \varnothing$ be the collection of selected concept indices.
2. **Ranking**: For every concept index $i \in M$, compute $\lambda_i = w_{ii}/t_{ii}$ and sort the index set M in ascending order according to $(\lambda_i)_{i \in M}$ values. The smaller the λ_i, the better the separation power of the concept.
3. **Selection**: For the sorted index set M, choose q concepts with smallest λ-values. Denote this set as M_q and write $M = M \setminus M_q$ and $N = N \cup M_q$. Construct cross-product matrices W_{ii} and T_{ii}, in such a way that they correspond to N selected concept indices.
4. **Evaluation**: Test Λ_1 and $\Lambda_{2,1}$. If the test based on $\Lambda_{2,1}$ indicates no remaining information, then stop.
5. **Update**: For the remaining indices $j \in M$, compute $\lambda_j = (w_{ii} - \boldsymbol{w}_{j1} W_{11}^{-1} \boldsymbol{w}_{1j}) / (t_{jj} - \boldsymbol{t}_{j1} T_{11}^{-1} \boldsymbol{t}_{1j})$. This step is carried out to remove the effect of the already selected variables. Then the execution moves to Step 2 and the process is repeated. In practice choosing a large q (i.e. several concepts are selected at once), leads to a quick termination of the algorithm. which is preferable for on-line use. That is, for most topics a single iteration should give sufficiently good results.
6. **Output**: The collection of selected concepts corresponding to the index set N.

3.3 Value function based classification

In the Wilks-VF framework, the utility or value of each document is obtained based on a combination of the individual attributes (i.e. concepts). Learning a filtering rule in this system is in essence equal to finding the optimal parameters for the user's value function. The process used in this paper is formalized as follows:

Definition 2. *Value Function: Let D denote the space of profiled documents, where each $d \in D$ is a vector of Wikipedia concepts. A value function representing the user's preference information is defined as mapping $V : D \to \mathbb{R}$, given by*

$$V(d) = \sum_{c \in C_N} \mu(c, d) w(c),$$

where $w(c) \in [-1, 1]$ denotes the weight of concept c and C_N is the set of selected concepts from the dimension reduction step. The function $\mu : C_N \times D \to \{0, 1\}$ determines the presence of a concept in the document by the rule

$$\mu(c, d) = \begin{cases} 1 \text{ if } d\text{-}rel(c, d) \geq \alpha \\ 0 \text{ otherwise} \end{cases}$$

where $d\text{-}rel(c, d) = \max_{\bar{c} \in d} c\text{-}rel(c, \bar{c})$ is a document-concept relatedness measure.

Let $\mathbb{D}^{(R)}$ and $\mathbb{D}^{(IR)}$ denote the set of relevant and irrelevant documents originally supplied by the user. The parameters of the value function are determined by solving the following linear optimization problem. Maximize ϵ subject to

$$V(d^{(R)}) - V(d^{(IR)}) \geq \epsilon$$
$$\forall d^{(R)} \in \mathbb{D}^{(R)}, d^{(IR)} \in \mathbb{D}^{(IR)}$$

A positive weight indicates a relevant concept and a negative an irrelevant one. When $\epsilon > 0$, the obtained value function is consistent with the user's preferences. Based on the Wilks-VF a simple document classification rule is obtained by choosing a suitable cutoff [1]. If a document's value is above the cutoff, it is considered relevant.

4 Experiments and Results

In this section, we present the results of the Wilks-VF method. The method has been tested on Reuters TREC-11 newswire documents, which is a collection of news stories from 1996 to 1997. The collection is divided into 100 subsets or topics. All documents belonging to a topic are classified as either relevant or irrelevant to the given topics. These topics are further partitioned into a training and an evaluation set. The purpose of the training set is to generate a search query, which is then applied onto the evaluation set in order to evaluate its performance.

The results from all 100 topics are reported together with five benchmark methods in Table 1. As benchmarks, we consider the following commonly applied feature selection techniques: Gain ratio [18], Kullback−Leibler divergence [18], Symmetric uncertainty [19], SVM based feature selection [20] and Relief [22] [23]. The performance is recorded in terms of precision and recall. F-Score is used to combine the two measures as: F-Score = (2 × Precision × Recall)/(Precision + Recall). The reported performance measures are calculated as averages over all topics. In the experiment we set $q = 10$.

Model	F-Score	Recall	Precision	Model	F-Score	Recall	Precision
GR	0.2785	0.2930	0.3902	SVM	0.3680	0.4073	0.4215
Symmetry	0.3235	0.3578	0.4110	Relief	0.3785	0.4568	0.4068
KLD	0.3391	0.3977	0.3984	Wilks-VF	0.3990	0.5184	0.4011

Table 1. Model comparison

As can be seen from the table, the Wilks-VF method is competitive in terms of F-Score. When searching for reasons, it appears that the performance differences are largely explained by recall levels. The recall for Wilks-VF is considerably better than other methods, as can be observed in Table 1. Differences across precision are however, smaller for different methods. Thisproposed by Cilibrasi and Vitanyi means that the advantage of Wilks-VF is its ability to retrieve relevant instances.

5 Conclusions

The paper discusses an important aspect in document classification, i.e. dimensionality reduction. Dimensionality reduction is ubiquitous in various fields and has been widely studied. In this paper, we have specialized a well known Wilks lambda procedure for document classification. The novelty introduced in the approach is a cluster based concept selection procedure which ensures that all the concepts which are significant for classification are selected. The dimensionality reduction procedure has been integrated with a recently suggested value function approach which makes the overall system computationally less expensive to an extent that the methodology can be developed for on-line usage. The empirical results computed on the Reuters TREC-11 corpus show that the Wilks-VF approach is efficient when compared with other widely used methods for dimensionality reduction.

References

1. Malo P., Sinha A., Wallenius J. & Korhonen P.: Concept-based Document Classification Using Wikipedia and Value Function. Journal of American Society of Information Science and Technology (2011) to appear
2. Rajapakse R. & Denham M.: Text retrieval with more realistic concept matching and reinforcement learning. Information Processing and Management (2006) vol. 42, 1260-1275
3. Kim H.: ONTOWEB: Implementing an ontology-based web retrieval system. Journal of American Society for Information Science and Technology (2005) vol. 56, no. 11. 1167-1176
4. Kim H.: Toward Video Semantic Search Based on a Structured Folksonomy: Journal of the American Society for Information Science and Technology (2011) 478-492
5. Rao C, Linear Statistical Inference. Wiley, NYC, NY (1973)
6. Pera M., Lund W. & Ng Y.-K.: A Sophisticated Library Search Strategy Using Folksonomies and Similary Matching. Journal of the American Society for Information Science and Technology (2009) vol. 60, 1392-1406

7. Malo P., Siitari P. & Sinha A.: Automated Query Learning with Wikipedia and Genetic Programming. Artificial Intelligence (2010) Conditionally accepted
8. Gabrilovich E. & Markovitch S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In Proc. IJCAI-07 (2007) 1606-1611
9. Malo P., Siitari P., Ahlgren O., Wallenius J. & Korhonen P.: Semantic Content Filtering with Wikipedia and Ontologies, 10th IEEE International Conference on Data Mining Workshops 2010, Los Alamitos, CA, USA: IEEE Computer Society (2010) 518-526
10. Ponzetto S. & Strube M.: Knowledge Derived From Wikipedia For Computing Semantic Relatedness. Journal of Artificial Intelligence Research (2007) vol. 30, 181-212
11. Milne D. & Witten I.: Learning to link with Wikipedia. Proc. CIKM, (2008)
12. Medelyan O., Milne D., Legg C. & Witten I.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies (2009) vol. 67, 716-754
13. Korhonen P., Moskowitz H. & Wallenius J.: A progressive algorithm for modeling and solving multiple-criteria decision problems, Operations Research (1986) vol. 34, no. 5, 726-731
14. Korhonen P., Moskowitz H., Salminen P. & Wallenius J.: Further developments and test of a progressive algorithm multiple-criteria decision making, Operations Research (1993) vol. 41, no. 6, 1033-1045
15. Deb K., Sinha A., Korhonen P., & Wallenius J.: An interactive evolutionary multi-objective optimization method based on progressively approximated value functions, IEEE Transactions on Evolutionary Computation (2010), vol. 14, no. 5, 723-739
16. Zionts S. & Wallenius J.: An interactive programming method for solving the multiple criteria problem. Management Science (1976) vol. 22, 656-663
17. Roy A., Mackin P., Wallenus J., Corner J., Keith M., Schmick G. & Arora H.: An interactive search method based on user preferences. Decision Analysis (2009) vol. 5 203-229
18. Abeel T., Van de Peer Y.& Saeys Y.: Java-ML: A Machine Learning Library. Journal of Machine Learning Research 10 (2009) 931-934
19. Yu L. & Liu H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning (2003) 856-863
20. Guyon I., Weston J., Barnhill S. & Vapnik V.: Gene selection for cancer classification using support vector machines. Machine Learning. (2002) 46:389-422
21. Burges C.: A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, Boston
22. Kira K. & Rendell L.: A Practical Approach to Feature Selection. Ninth International Workshop on Machine Learning (1992) 249-256
23. Robnik-Sikonja M. & Kononenko I.: An adaptation of Relief for attribute estimation in regression. Fourteenth International Conference on Machine Learning (1997) 296-304
24. Harris E.: Information Gain Versus Gain Ratio: A Study of Split Method Biases. The MITRE Corporation (2001)
25. Fodor I.: A Survey of Dimension Reduction Techniques. U.S. Department of Energy (2002)
26. Cilibrasi R. & Vitnyi P. M. B.: The Google Similarity Distance. IEEE Trans. Knowl. Data Eng. (2007) 19(3): 370-383
27. Friedman H. P. & Rubin J.: On Some Invariant Criteria for Grouping Data. Journal of the American Statistical Association (1967) vol. 62, no. 320 , 1159-1178