LATENT SEMANTIC INDEXING USING MULTIRESOLUTION ANALYSIS

Tareq Jaber[1], *Abbes Amira*[2], *Peter Milligan*[3]

[1]Faculty of Computing and Information Technology
King Abdulaziz University - North Jeddah Branch, KSA
tjaber1@kau.edu.sa
[2]Nanotechnology and Integrated Bio-Engineering Centre (NIBEC), Faculty of Computing and Engineering
University of Ulster, Jordanstown campus, Antrim, BT37 0QB, UK
a.amira@ulster.ac.uk
[3]School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, Belfast, BT7 1NN,UK
p.milligan@qub.ac.uk

ABSTRACT

Latent semantic indexing (LSI) is commonly used to match queries to documents in information retrieval (IR) applications. It has been shown to improve the retrieval performance, as it can deal with synonymy and polysemy problems. This paper proposes a hybrid approach which can improve result accuracy significantly. Evaluation of the approach based on using the Haar wavelet transform (HWT) as a preprocessing step for the singular value decomposition (SVD) in the LSI system is presented, using $Donoho's$ thresholding with the transformation in HWT. Furthermore, the effect of different levels of decomposition in the HWT process is investigated. The experimental results presented in the paper confirm a significant improvement in performance by applying the HWT as a preprocessing step using $Donoho's$ thresholding.

I. INTRODUCTION

As the amount of information stored electronically increases, so does the difficultly in searching it. The field of information retrieval (IR) examines the process of extracting relevant information from a dataset based on a user's query [1]. Latent semantic indexing (LSI) is a technique used for intelligent IR. It can be used as an alternative to the traditional keyword matching IR and is attractive in this respect due to its ability to overcome problems with synonymy and polysemy [1]. Traditionally LSI is implemented in several stages [2]. The first stage is to preprocess the database of documents, by removing all punctuation and "stop words" such as the, as, and etc, those without distinctive semantic meaning, from a document. A term document matrix (TDM) is then generated which represents the relationship between the documents in the database and the words that appear in them. Then the TDM is decomposed. The original decomposition algorithm proposed by Berry [1] et al, and by far the most widely used, is the singular value decomposition (SVD) [3]. The decomposition is used to remove noise (sparseness) from the matrix and reduce the dimensionality of the TDM, in order to ascertain the semantic relationship among terms and documents in an attempt to overcome the problems of polysemy and synonymy. Finally, the document set is compared with the query and the documents which are closest to the user's query are returned. In Unitary Operators on the Document Space [4], Hoenkamp asserts the fundamental property of the SVD is its unitary nature. And the use of Haar wavelet transform (HWT), as an alternative that shares this unitary property at much reduced computational cost, has been suggested, and this research presents some promising initial results. Further the idea of the TDM as a gray scale image has also been postulated, and the equivalence of using the HWT to remove lexical noise and using the HWT to remove noise from an image has been discussed [4]. The aim of the research presented in this paper, and continuing on the research work in [5], is to develop a new approach to the LSI process based on the possibility of using image processing techniques in text document retrieval. In particular, the effect of using the HWT as a pre-processing step to the SVD is studied. Moreover, attention is paid to the effect of different levels of decomposition and threshold techniques used in the HWT. A range of parameters and performance metrics, including accuracy or precision (number of relevant documents returned), computation time, and threshold value or dimensions retained, are used to evaluate the proposed LSI system. The paper is organized as follows. The proposed hybrid approach is presented in section 2. This gives an overview of our proposed LSI system and the processes involved. Section 3 presents the results of the new method as a series of hypothesis which are evaluated by comparing to standard baseline systems. Concluding remarks are given in section 4.

II. THE PROPOSED HYBRID METHOD

In this section the proposed framework with its different components are illustrated. To enable evaluation of the modified approach, the method is applied to four sample databases: Memos database [1][2], Cochrane database [6], eBooks database [7], and Reuters database [3].

II-A. The HWT

The HWT is a series analogous to the Fourier expansion that is often used in image processing [8]. HWT decomposition works on an averaging and differencing process [8][9]. In image processing, the transform can be used to remove noise from an image [10]. An image is transformed using HWT, and then a thresholding function, at a certain threshold value, is applied to remove the noise from the image; typically a cleaner image results when the image is reconstructed after thresholding. $Donoho's$ thresholding algorithm is used in this paper [11]. This algorithm generates the threshold value by the given equations:

$$\lambda = \gamma\sigma\sqrt{\frac{2\log(n)}{n}} \tag{1}$$

where λ is a threshold value, n is a number of sample data, σ is a noise standard deviation and γ is a constant. It can be seen that this threshold depends simply on the value of sample data. Then the hard-thresholding function, or the soft-thresholding function [12], is applied to threshold the image. The paper will present the results

of applying each thresholding model in the new approach. For our present purposes, the TDM can be considered as a gray scale image, usually a binary image (sparse TDM with 0, 1 probabilities). By applying the HWT and a threshold to the TDM, we can also remove "noise" from our image, in this case we argue that this represents the removal of lexical noise.

II-B. HWT/SVD based hybrid approach

A commonly used approach in image processing is to combine different techniques in order to improve noise reduction. The comparison of the TDM to a gray scale image invites a similar technique. The system allows SVD and HWT techniques to be combined, as shown in Fig. 1, to investigate their combined effect on the TDM and the quality of the results, in terms of $precision$ and $recall$.

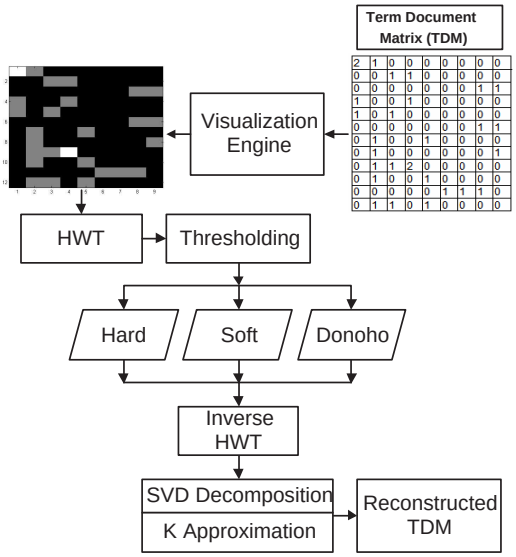

Fig. 1. The hybrid method

II-C. Lexical Noise Analysis

In the LSI process, once preprocessing is complete, the TDM is constructed. Most values in the matrix are zero, as only a subset of keywords appears in any given document. It is interesting to see the relationship of terms across documents; therefore words that appear only in one document add no information to this relationship. Similarly, words that appear in all documents are considered meaningless due to their ubiquity. Fig. 2 shows images generated by visualizing the TDMs after applying the HWT/SVD hybrid method. It is worth noting that visualizing the TDM as an image enables large datasets to be examined and analyzed more easily [13]. If the images are examined, the white dots represent the data or non-zero values. When dots are close to each other, forming a cluster, it is possible, by looking at appropriate columns and rows, to say that there is a relationship between these documents because they reference about the same concepts.

III. RESULTS AND ANALYSIS

This section compares the proposed hybrid techniques with the standard LSI approach. The search is applied to the different databases. There are several different measures for evaluating the performance of information retrieval systems. The most common properties that are widely accepted by the research community are $recall$ and $precision$ [14]. $Precision$ is the fraction of the documents retrieved that are relevant to the user's query, and $recall$ is the fraction of the documents that are relevant to the query and successfully retrieved [5].

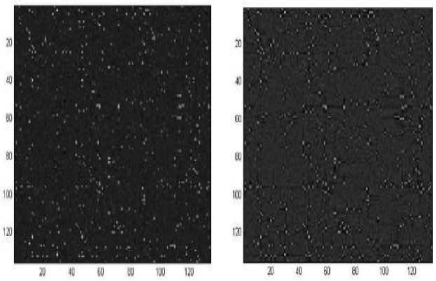

Fig. 2. Left: SVD Decomposition of Cochrane database with K= 40 Right: HWT/SVD of Cochrane Database with threshold = 0.04 and K=40. The SVD image is slightly brighter, and has more dark pixels which represent zeros. For HWT image, the great deal of sparseness (dark pixels) has been reduced and values are redistributed more evenly

Both $recall$ and $precision$ are needed for measuring issues in the IR. It is common to achieve recall of 100% by returning all relevant documents in response to any query, therefore recall alone is not enough. One needs to measure the number of irrelevant documents to determine the precision or accuracy of the results returned. On the other hand, precision of 100% can also be achieved in many cases by returning only relevant results, but again, one needs to count all the relevant documents in the database to measure the recall.

III-A. HWT-SVD LSI

In this section a number of searches are performed on the sample databases to compare the basic LSI-SVD approach with the proposed hybrid technique. (The standard SVD when mentioned in the work indicates the standard LSI-SVD system).

- **Cochrane Database:** Searching for "rheumatoid arthritis" and "smoking and heart disease"

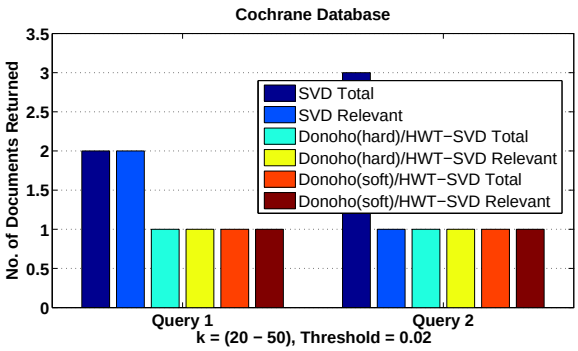

Fig. 3. LSI search results for Cochrane database

For the first query, as shown in Fig. 3, the standard SVD returns one more result than the hybrid approach. At the second query, the standard method has a low precision value, by returning three results, two of which are irrelevant to the query. The HWT-SVD method returns only related documents demonstrating greater accuracy or precision in matching queries, and hence outperforms the standard method.

- **eBooks Database:** Searching for ”plastics engineering”,”xml transformations”,”health and safety” and ”advanced java programming”

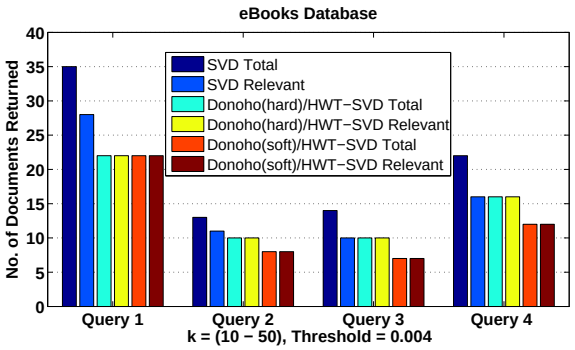

Fig. 4. LSI search results for eBooks database

In Fig. 4, for the first query as shown before, the standard SVD returns all the existing relevant results, and outperforms the HWT-SVD approach with recall of 100% and 79% for the new novel approach with both thresholding. However, and as shown in the figure, the standard SVD produces seven irrelevant documents, while both thresholding schemes, with the new approach, return 22 documents, all of which are relevant with precision of 100% and 80% for the standard SVD. In the second query, with precision of 100%, the Donoho(hard)/HWT-SVD approach returns one less relevant result, and the Donoho(soft)/HWT-SVD returns three less relevant results than the standard SVD, while the standard SVD produces two extra unrelated documents. For the third query, the standard method returns four unrelated extra documents resulting in a precision of 71%, while the hybrid method with the hard function obviously improves the accuracy level with precision of 100%. A recall of 100% for the both methods is achieved as they retrieve all the related results in the database. The new method with the soft thresholding performs not very well and returns three less related documents with recall of 70%. For the last query, Donoho(hard)/HWT-SVD again performs very well by eliminating six additional unrelated documents returned by the standard method, resulting in a considerable improvement for the precision volume with recall and precision of 100%. The Donoho(soft)/HWT-SVD returns four less relevant documents than the hard function.

- **Reuters Database:** Searching for ”japan”,”bank”,”money market” and ”sales tax”

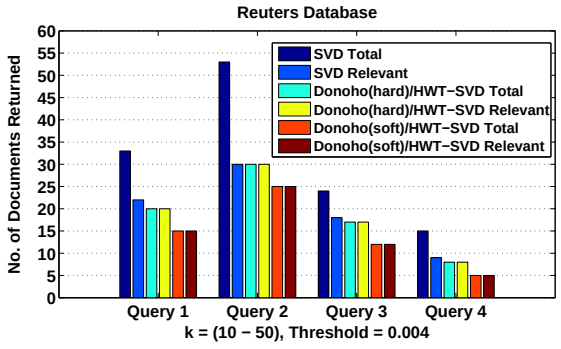

Fig. 5. LSI search results for Reuters Database

At the first query for the search in Fig. 5, the standard method as shown before performs inefficiently in the precision action, with a large volume of irrelevant results returned, causing a low value of precision (67%). Reasonable results are returned by the new hybrid approach with the hard function. The new method returns only two less relevant results and keeps only the relevant documents with recall of 91% and precision of 100%. A less efficient performance can be noticed in the new approach with the soft function. Although precision of 100% is achieved, this method misses seven related documents, which decreases the recall to 68%. Excellent results are obtained by the Donoho(hard)/HWT-SVD for the second query, the method returns all the existing relevant documents to the query without any irrelevant results, with recall and precision of 100% obtained. As shown in the previous section, a considerable volume of irrelevant documents are returned by the standard SVD with precision of 53%, and thus, shows a poor accuracy level. As was the case in the previous query the Donoho(soft)/HWT-SVD again performs less well than the hard one. All the documents returned are relevant but it produces a lower volume of relevant results, with a recall of 83% and a precision of 100% achieved. The standard SVD at the third and fourth queries keeps showing a lower level of accuracy for the results returned compared to the hybrid novel approach, with both thresholding methods. A negligible lower number of relevant results is returned by Donoho(hard)/HWT-SVD, resulting in a slightly lower recall value. As noticed, in most of the cases, the Soft/HWT-SVD misses more relevant results and as a consequence the recall value decreases.

Table I. Decomposition Algorithms Computation Time (seconds) and Accuracy

Database	SVD	$Donoho's$/HWT-SVD
Memos	0.11	0.245
Cochrane	0.562	0.977
eBooks	16.562	19.765
Reuters	27.125	33.314
Accuracy	66%	100%

Table I provides an overview of the computation time in seconds and the accuracy action for the different decomposition algorithms. As the size of the matrix grows, the amount of extra processing time required to implement HWT preprocessing into the LSI system becomes negligible in comparison to the overall time required, especially with the improvement in accuracy of the returned results.

III-B. Multilevel Decomposition Analysis

This section investigates the influence of the decomposition level on the search results, by applying the search at the best k (the dimension reduction in the SVD algorithm for which the best results are obtained) and threshold value at different levels of decomposition. The results in the previous section show that the Donoho(hard)/HWT-SVD in many cases performs better than the Donoho(soft)/HWT-SVD. Consequently the threshold function used in this investigation is the hard function.

Fig. 6 shows that at range of decomposition levels (4-8) the method keeps returning one relevant result, while no results are returned at the levels less than four.

In Fig. 7 the results show that at the decomposition level ten, which is the full decomposition level, the maximum number of results is obtained. As the level of decomposition decreases, the number of results returned decreases as well.

Fig. 8 presents similar results as shown before, no documents are produced for level less than five, while only one related document is returned at this level. A far larger number of relevant results is obtained at the higher levels (6-10).

III-C. Analysis

In this research we argue that low frequencies (0 values in a typical TDM) represent lexical noise (or unrelated documents) and, consequently, the deletion of these values will not affect the

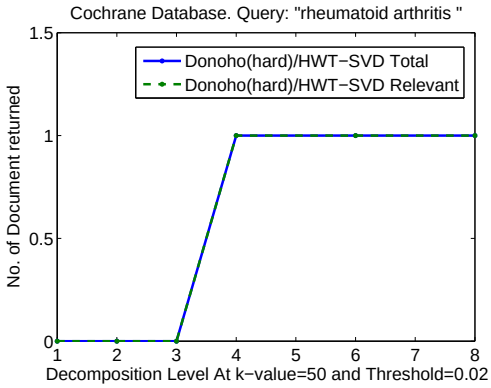

Fig. 6. Multiresolution analysis for Cochrane database

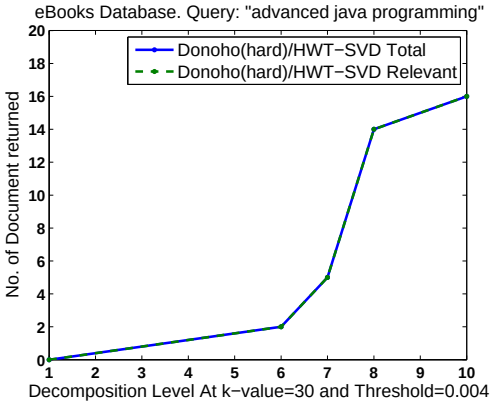

Fig. 7. Multiresolution analysis for eBooks database

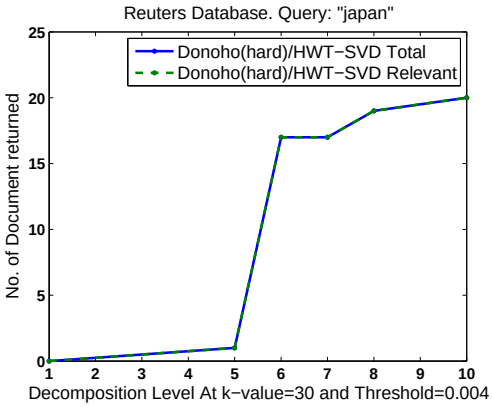

Fig. 8. Multiresolution analysis for Reuters database

structure of the TDM. The elimination of this noise is effected by the application of the wavelet transform which separates low and high frequencies and the subsequent use of the threshold function which removes low values from the TDM. Finally, the use of the inverse HWT generates a new TDM. For the hybrid HWT-SVD, the results show that by adding the HWT as a pre-processing step the precision is improved by approximately 34%. The HWT-SVD, in most cases, does not produce any irrelevant documents, and returns the same number of relevant documents that are returned by the standard SVD approach. The HWT-SVD uses $Donoho's$ thresholding to generate the threshold value, and then thresholds the TDM using two thresholding functions. The results show that the Donoho(hard)/HWT-SVD performs clearly better than the Donoho(soft)/HWT-SVD in terms of the number of relevant results returned. In the investigation of different levels of decomposition, the results show that at the level of full decomposition the best results are obtained. It is beneficial to note that the precise action of the pre-processing step depends on the value of k used for the SVD and the threshold value used in HWT, (that are chosen by testing the methods at range of k and threshold values), but for most optimal cases the results show that adding HWT pre-processing can improve the precision of the documents returned.

IV. CONCLUSIONS

A new hybrid modified approach to LSI for effective use in IR has been presented in this paper. Investigation of different approaches for LSI has confirmed that the SVD is the most powerful decomposition algorithm in the LSI process in terms of the number of documents returned. The results of the investigation for the HWT as a preprocessing step, prior to the SVD in the LSI process, shows good results, the preprocessing step tends to remove irrelevant documents from the documents returned, causing enhancement of the accuracy of the results returned. The multiresolution analysis shows that HWT performs better at the full decomposition. It also offers the possibility of other combined approaches.

V. REFERENCES

[1] M. Berry, S. Dumais, and G. OBrien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, vol. 37, pp. 573 – 595, 1995.

[2] M. Bell and N. Degani, "Latent semantic indexing, parallel svd and its applications," *Proceedings of ALGORITMY 2002*, pp. 113–120, 2002.

[3] C. Fox, "Lexical analysis and stoplists. in information retrieval - data structures & algorithm," *Prentice-Hall*, pp. 102–130, 1992.

[4] E. Hoenkamp, "Unitary operators on the document space source," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 314–320, 2003.

[5] T. Jaber, A. Amira, and P. Milligan, "Performance evaluation of dct and wavelet transform for lsi," *IEEE International Symposium on Circuits and Systems (ISCAS). Seattle, USA*, 2008.

[6] Cochrane, "Url: http://www.cochrane.org," 2005.

[7] eBooks, "Url: http://www.library.qub.ac.uk," 2005.

[8] A. Amira and P. Farrell, "An automatic face recognition system based on wavelet transforms," *Proceedings of the IEEE International Conference on Circuits and Systems*, pp. 6252–6255, 2005.

[9] M. W. Berry, Z. Drmavc, and E. R. Jessup, "Matrices, vector spaces, and information retrieval," *SIAM Review*, vol. 41, pp. 335–362, 1999.

[10] I. Delakis, O. Hammad, and R. I. Kitney, "Wavelet-based denoising algorithm for images acquired with parallel magnetic resonance imaging (mri)," *Physics in Medicine and Biology*, vol. 52, pp. 3741–3751, 2007.

[11] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transaction on Information Theory*, vol. 41, pp. 613–627, 1995.

[12] B. Yoon and P. P. Vaidyanathan, "Wavelet-based denoising by customized thresholding," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 925–928, 2004.

[13] T. Jaber, A. Amira, and P. Milligan, "A novel approach for lexical noise analysis and measurement in intelligent information retrieval," *Proceedings of IEEE International Conference*

on Pattern Recognition ICPR, Hong Kong, vol. 3, pp. 370–373, 2006.

[14] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Data Engineering Bulletin*, vol. 24, pp. 35–43, 2001.