AN APPROACH TO DETECT ILLEGAL SIMILARITY IN RESEARCH LITERATURE USING LATENT SEMANTIC INDEXING

Muna Alsallal, Rahat Iqbal

Abstract: Research suggests that there are an increasing number of illegal similarities within research literature. As part of our research we are investigating the application of an information retrieval technique, Latent Semantic Indexing, to derive semantic information from text files. In this paper, we present an integrated framework for enhancing the automatic detection for illegal similarity texts, steering the area of Latent Semantic Index (LSI) and highlighting its ability to unmask the latent relationship between texts in order to detect illegal similarity. We have conducted an experiment to investigate the efficiency of a dimensionality reduction parameter as the core for LSI technique, the experiments designed to establish the highest amount of re-occurrence for given values and also the distribution for given values of the dimensionality reduction parameter k in Latent Semantic Indexing. The results so far are promising.

Introduction

The adopting of others' ideas and manipulation of the findings have overloaded research literature which makes it very difficult to optimize results of information retrieval and data mining systems to match user information needs. Some on-going studies suggest excuses for such a risky phenomenon stating that researchers may not have sufficient time to track their own ideas, and publishers may not be well-equipped to check whether the contributions and results come from original research (Alzahrany et el, 2012). Misconduct in research is becoming increasingly more sophisticated, including incidence of duplication of publications and illegal reuse of others' research ideas, contributions and findings, (Hennessey et el, 2012). There is scope for future research to address these problems by exploring use of algorithms to digitally analyse essential literature sources.

The Research Problem

The problem of illegal similarity in research literature is currently solved by a variety of different algorithms which function by comparing new manuscripts to the already published texts contained by a large library. This comparison works by means of comparing significant key words, phrases which are statistically improbable, and by computing a measure of similarity on the basis of aligning different sentences (Nayak, 2009), If this information exceeds a particular threshold, it may then alert the user to the possibility of illegal similarity.

However, there are a number of factors which need to be kept in consideration when deciding upon the effectiveness of an algorithm. For example, Lewis et al (2006) argues that the effectiveness of the algorithm will depend on some factors such as the number of databases which are served, the extent to which it is compatible with the journal

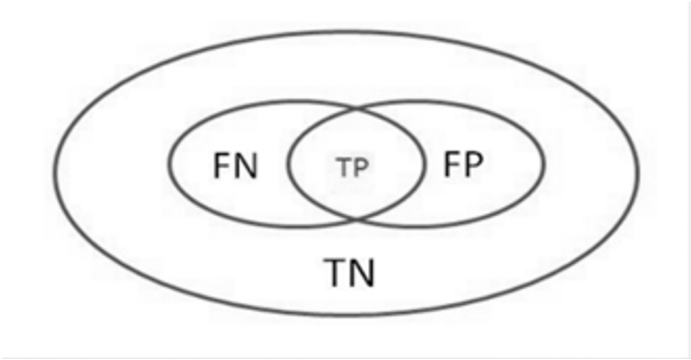

Figure 1. Venn diagram shows the current detection systems performance

manuscript submission system, the user interface (in particular, the extent to which the results are presented in a manner which is meaningful and easily assimilated) and the security of data. This is expanded upon by Garner(2011) who argue that the effectiveness of the algorithm at identifying illegal similarity is closely related to the specificity and the sensitivity of the search algorithm. It is therefore important to ascertain the way in which the search algorithm works, and what the false negative (documents that share suspicious text but not retrieved as suspicious) and the false positive (documents that is retrieved as suspicious but in point of fact it is innocent) rates of the algorithm are. Figure1 shows the current detection systems performance. Furthermore, for ease of use, it is suggested that an effective algorithm should provide the user with threshold settings in order to enable them to minimise false negatives and false positives and to prioritise different manuscript sections where it is possible to tolerate differing levels of similarity (Long, Errami & George, 2009).

Contribution

The contributions of our work are threefold: Section 2 identifies the concept of plagiarism detection systems and it's privacy within research environment. Section 3 simplifies the complicated concept of LSI which helps the readers from different backgrounds to understand the state of art of this pure mathematical method. In section 4 the proposed framework based on intelligent techniques and algorithms work together to detect the new added articles semantically and syntactically and define the old ones as well, in addition our framework has beneficiary of PPR (post publication review) technique which will give the user a chance to participate in enhancing the performance of the detection process, put this system of the interest of open access articles users and finally we are working to build a robust system to detect different kinds of illegal text reuse but keep consideration on preventing the loss of knowledge.

Illegal Similarity Detection Approaches

Recent anti-plagiarism systems can identify only word-for-word similarity and only some incidence of it (Eissen et.el, 2007) and do not cover the ideas adoption or fake results. On-going research in this area classifies plagiarism detection systems into two types:

- Extrinsic plagiarism detection systems which deal with declared illegal text similarity as most of algorithms work using a technique to compare word for word to discover illegal similarities syntactically and some algorithms attempt to detect illegal similarity semantically. However these systems are gaining good results such as Turnitin which is used widely as a detection system for student essays, but this tool still faces difficulties in detecting some kinds of text manipulation. Approaches used to detect extrinsic illegal similarity required a reference collection of likely original texts.
- Intrinsic plagiarism detection systems work to unmask undeclared illegal text similarity, such as ideas adoption or fake results. In intrinsic detection approach, researchers attempt to simulate the human judgment in discovering the intelligent changes in some texts although they do not have a reference library in their memory. Stylometric feature extraction is one of the interesting research approaches that will be used in our proposed framework to enhance the detection process.
- English language becomes the crucial language for publishing stakeholders, that makes reuse some words or phrases in order improve the language and writing style a common practice for international researchers to deliver their ideas and findings in an acceptable style. However most current plagiarism detection identify the above practice as an illegal despite the originality of ideas and findings (Mason, 2009).

Latent Semantic Indexing (LSI)

LSI is mainly used as an information retrieval technique, which is generally based on the spectral analysis of the term document matrix (Edmunds 1997). It essentially captures the hidden structure within an information retrieval method using techniques from linear algebra (Veling and Van der Weerd 1999). What generally occurs is that there are vectors that represent the documents or text, which are projected in a low dimensional space which is then obtained by a singular value decomposition of the terms (Dhillon and Modha, 2001). The most interesting feature of LSI is language independent

How LSI in our Proposed Framework Works

The procedure has been carried out by implementing five steps:

1. Pre-processing by eliminating sole and popular terms from the corpus and activate stop word list which contains conjunctions, prepositions..etc
2. The functionality of LSI starts by transforming the pre-processed corpus into an m x n matrix in which its column represent documents and its rows represent terms (Berry et. al, 1993)

3. Apply term weighting scheme to eliminate the non-relevant terms as local weighting shows the term frequency in document while global weighting shows the term frequency in the corpus.
4. Apply SVD method to reduce the corpus into an intensive corpus containing only significant terms from the original by decomposing the original matrix into 3 concentrated matrices (Berry et. al, 2005).
5. Apply Dimensionality Reduction technique in order to truncate the three matrices that resulted from step 4 into k-dimensions by selecting the first k columns from each matrix and assign zero for the rest values.

Effective Parameters

Previous manuscripts have shown promising results while using latent semantic indexing to simulate the human ability to detect illegal similarity. The methods that have been conducted using LSI are quite different and resulted in putting huge obstacles to identify which parameters will lead to success or failure (Haley et al., 2005, 2007). We can summarize the most important parameters that may affect the performance of LSI into four groups:

1. Weighting schemes, which represent the importance of the term in document that so-called local term weight and global term weight which represents the importance of a specific term in the corpus. Wild et al. (2005) advised to use the formula (1) to calculate the local term weighting as they had achieved good results.

$$Log \qquad l_{ij} = \log(tf_{ij} + 1) \tag{1}$$

where tf_{ij} is the number of times that term i appears in document j.
For global weights, formula (2) is the common global weighting formula has been used. However (Harley et al. 2005) had achieved a very satisfactory output by using formula (3)

$$IDF \qquad g_i = \log_2(n/df_i) + 1 \tag{2}$$

where df_i is the number of documents where the term i appears

$$Entropy \qquad g_i = 1 + j(p_{ij}\log(p_{ij})/\log(n)) \tag{3}$$

where tf_{ij} is the number of times that term$_i$ appears in documentj; gf_i is the total number of times term i appears in collection; n is the number of documents in the collection.
2. Dimensionality Reduction: Many researchers elucidate that dimensionality reduction represent the mathematical core for LSI so it is valuable to investigate this parameter to establish the usefulness of LSI method. Since the idea of LSI depended on establishing the term-document matrix to represent the text and using SVD; the mathematical method to truncate the original matrix into k most relevant dimensions in order to eliminate the non-relevant documents (Zeimpekis

& Gallopoulos, 2005). Far from mathematical language, DR reduces large datasets to intense datasets containing only related data from the original data.

3. Similarity Measures: Many measure correlations have been used to measure the distance between two vectors, for instance the common similarity measure is cosine measure which practically widespread for plagiarism detection. The automatic detection systems are using the cosine measure between the suspicious document vector and relevant documents.

$$\cos(VS1; VS2) = \sum_{i=1}^{k} (VS1_i \cdot VS2_i) \frac{VS1_i * VS2_i}{|VS1| * |VS2|} \tag{4}$$

where $VS1$ is the vector representing the first suspicious file, $VS2$ is the vector representing the second suspicious file and k is the number of dimensions.

4. Pseudo documents: Pseudo documents are constructed to represent the content of documents in semantic spaces in order to extract the underlying relationship between terms. We can construct pseudo documents using two popular methods:
   - Vector Sum: Extract the meaning of documents via summation of word vectors of the document.
   - Folding-in procedure: Constructing the pseudo document via folding an extra document into k-dimensional vector space.

Outline of Framework

We propose a framework in order to increase the effectiveness in which illegal similarity in research literature can be identified. The proposed framework consists of group of components working together to enhance the detection task.

The framework which is outlined above is a proposal for automatic illegal similarity detection (ISDT) which fulfills the task of detecting illegal similarity semantically to declare the latent relationships between terms. It consists of the following components:

- Latent semantic indexing (LSI) which fulfills the role of detecting the incidence of semantic similarity between different texts.
- Modified Stylometry algorithm which fulfills the role of detecting the similarity between different texts depending on extracting the writing style features.
- Modulator which works by calculating the extent of similarity between two separate texts based on a given threshold value. Once this rating has been calculated, this information then entered onto the global similarity index.
- Global similarity index. It works as a management component that stocks the suspicious detected files in the master list of detected
- Medline index which is the principal index pertaining to the research literature
- A keyword based search tool which works by retrieving all of the documents which are relevant to a particular query.
- Normal User interface which consists of a search system interface which is employed for the searching of documents by users.

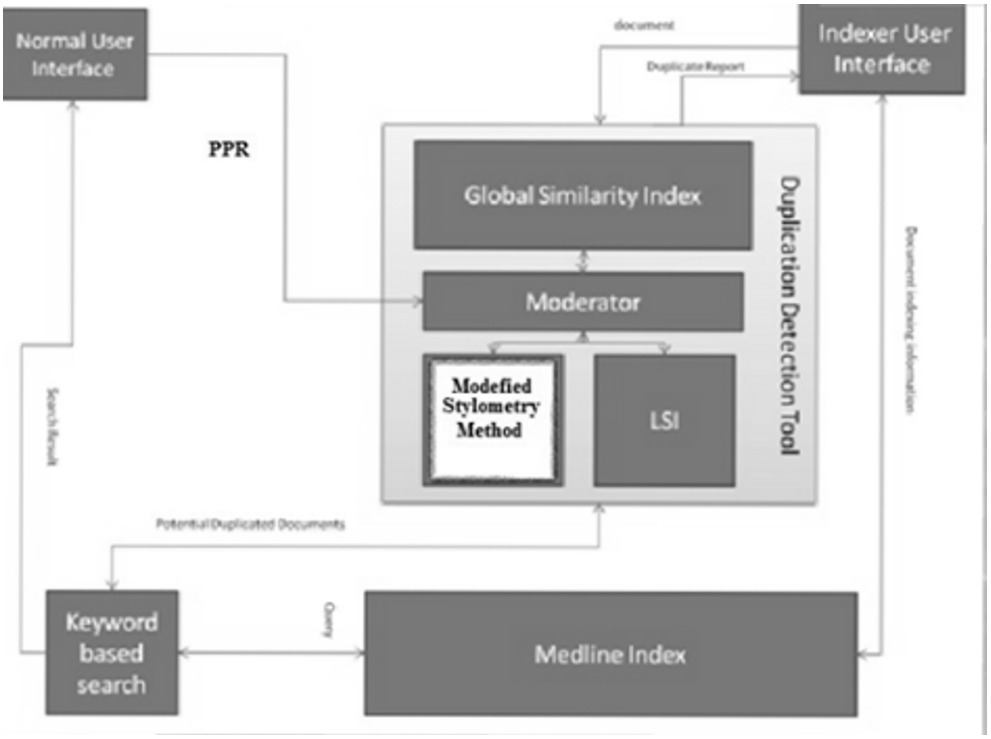

Figure 2. The Proposed Framework

- Post Publication Review (PPR) technique which will be used to get the user feedback on the results from ISDT. This opinion will be weighted and entered into moderator calculations.
- Indexer user interface which is used by individuals who add information to the Medline Index.
- PPR Technique: the user receives the results of the search together with a list of associated optional contextually alike documents. At this point, the user will be asked to manually assess the documents which have been retrieved to determine the potential identical documents in order to ascertain whether the level of similarity between both documents is sufficiently high; this manual assessment so-called post publication review technique (PPR) can be used as an important factor to enhance the detection process. PPR factor will in-turn be employed in moderator calculations. The identical similarity process will be iterated taking into its account the value of PPR factor.

Method

According to Lewis et al (2006), the key distinguishes features of LSI is the fact that the meaning of terms is derived by means of an approximation of the structure of term

Table 1
Deerwester (Deerwester et. al, 1990) co-occurrence / Illegal similarity summary

Degree	Number of Occurrences / Duplicates
1	30
2	22
3	12
4	3

usage in different documents using the aforementioned decomposition algorithm. We conducted two stages experiments the first stage without using LSI method and the second stage with using LSI method.

Experiment Details

The experiments were carried out to investigate the significance of using dimensionality reduction parameter to highlight the bridge terms and its importance in unmasking the hidden relationships between two terms that don't coincide together but instead they coincide with a joint term so-called a bridge term. We conducted the experiments using MED collection for both of experiments to investigate the performance of LSI comparing with Find-Transitivity program.

MED Collection

The collection has been used to find singular values and vectors for k = 50, 100, 200 and 300. The singular values and vectors were then be an input for find-transitivity program; this identifies the level of transitivity for each pair of identical pairs (Swanson, 1991). Further work was carried out which developed a term by term duplicate matrix. The find-transitivity program takes an input in the form of vectors and singular values produced by the SVD program.

Results and Discussion

A summary of the results found by using the find-transitivity program is shown as follows:

From stage (1), four was the highest order that was observed in the case of the find-transitivity program for the MED data.

From stage (2) the following table shows the results up to third order co-occurrence, which was examined.

The table above shows the results for the MED inputs to degree level 3. There were no significant changes in term of **k**, however it was noted that there was a slight upwardly trend, when **k** increases. The MED data reveals that there are a total of 1,110,491 pairs of terms that can be accessed or connected directly. A total of 15,869,045 pairs of terms are connected when one intermediate node is allowed, whilst there are 17,829 pairs of terms that can be connected with two intermediate nodes. Each of the terms within

Table 2
Results from MED Testing

	Degree 1	Degree 2	Degree 3
k = 50	1,110,483	15,866,518	17,817
k = 100	1,110,485	15,867,200	17,819
k = 200	1,110,491	15,867,595	17,824
k = 300	1,110,491	15,867,595	17,824

the MED data can be connected to one another with two intermediate nodes or less between terms.

It is clear from the evidence shown that transitivity within the co-occurrence relationship plays a key role in ensuring the effectiveness of any system that uses tools such as information retrieval, computational linguistics and textual data applications. Results from both types of experiments carried out have concluded that there is a means of improving the illegal similarity detection within a system in a successful manner, there is grounds for further investigations to combine the Dimensionality Reduction (DR) parameter to enhance the results as DR perform well when connects terms through bridge terms nevertheless in term of k different values doesn't give a motivating pointer. In addition the optimum of parameters combination will increase the effectiveness of LSI method to detect illegal similarity between research papers. Further experiments using different parameters will be carried out in order to achieve the goal of the proposed framework.

References

Berry, Michael, Theresa Do, Gavin O'Brien, Vijay Krishna, and Sowmini Varadhan. 1993. SVDPACKC (Version 1.0) User's Guide. April 1993.

Berry, MW. Pulatova, SA. & Steward, GW. 2005. Computing sparse reduced-rank approximations to sparse matrices, *ACM TOMS* 31, no. 2.

Cosma, G. & Joy, M., 2006. Source code plagiarism: a UK academic perspective, *Research Report*, No. 422, Dept of Computer Science, Univ of Warwick, Coventry

Cosma, G., 2008. An approach to source code plagiarism detection and investigation using latent semantic analysis, Thesis, University of Warwick, available from `http://www.dcs.warwick.ac.uk/report/pdfs/cs-rr-440.pdf` accessed 30 January 2013

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6): 391–407.

Dhillon, IS. & Modha, DM. 2001. "Concept Decompositions for Large Sparse Text Data using Clustering", Machine Learning, 42:1, pp. 143–175.

Edmonds, Philip. 1997. Choosing the Word Most Typical in Context Using a Lexical Co-occurrence Network. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Pages 507–509.

Errami, M. & Garner, H. R., 2008. A tale of two citations, *Nature*, vol. 452, pp. 397–9

Garner, H. R., 2011. Combating unethical publications with plagiarism detection services, *Urologic Oncology: Seminars and Original Investigations,* vol. 29, pp. 95–99

Gollogly, L. & Momen, H., 2006. Ethical dilemmas in scientific publication: pitfalls and solutions for editors, *Rev Saude Publica,* vol. 40, pp. 24–9

Haley, D. T., Thomas, P., Petre, P., & De Roeck, A. (2007). Seeing the whole picture: Comparing computer assisted assessment systems using LSA-based systems as an example. Technical Report Number 2007/07.

Hennessey, K. Williams, A. Afshar, A. MacNeily, A. (2012) Duplicate publications: A sample of redundancy. Journal of Urology, 6(3), pp. 177–180

T. C. Hoad and J. Zobel. Methods for Identifying Versioned and Plagiarised Documents. JASIST, 54(3): 203–215, 2003.

Lewis, J., Ossowski, S. & Hicks, J., 2006. Text similarity: an alternative way to search MEDLINE, *Bioinformatics,* vol. 22, pp. 2298–304

Long, T. C., Errami, E. & George, A. C., 2009. Scientific integrity: responding to possible plagiarism, *Science,* vol. 323, pp. 1293–4

Mason PR. Plagiarism in scientific publications. J Infect Dev Ctries. 2009;3:1–4.

Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007). Plagiarism detection without reference collections. Advances in Data Analysis (pp. 359–366) Springer.

Moussiades, L. M. & Vakali, A., 2005. PDetect: A clustering approach for detecting plagiarism in source code data sets, *The Computer Journal,* vol. 48, pp. 651–661.

Mozgovoy, M., 2007. Enhancing computer-aided plagiarism detection, Dissertation, Department of Computer Science, Univ of Joensuu.

Nayak, B. K., 2009. Author's misconduct inviting risk: duplicate publication, *Indian Journal of Opthalmology*, Nov–Dec, 57(6), pp. 417–418.

Sch□tze, Hinrich. 1998. Automatic Word Sense Disambiguation. Computational Linguistics, Volume 24, number 1.

Swanson, DR. 1991. Complementary Structures in disjoint science literatures. In A. Bookstein, et al (Eds), SIGIR91: Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. pp 280–289.

Veling, Anne and Peter van der Weerd. 1999. Conceptual grouping in word co-occurrence networks. Proceedings of the IJCAI '99. Volume 2. Pages 694–699.

Zeimpekis, D. & Gallopoulos, E., 2005. TMG: A MATLAB Toolbox for generating term-document matrices from text collections, in *Grouping multidimensional data: recent advances in clustering,* J Kogan & C. Nicholas ed. New York, USA: Springer Publishers.

Acknowledgements

Personally, I would like to thank Mrs Irene Glendinning; my line manager for encouraging me all the time in addition to her continues support. Without her care; was not easy to overcome hard times. I truly appreciated all her time and advices.

I would like to thank Dr. Rahat Iqbal and appreciated his guidance, support and willingness to take time to discuss my research.

Authors

Muna Alsallal, aa6095@coventry.ac.uk, Rahat Iqbal r.iqbal@coventry.ac.uk, Coventry University, United Kingdom

About authors

Muna Alsallal—BEng in Computer Hardware and Architecture Msc in IT and Application
PhD Candidate in Computing and Engineering Coventry University.

Dr Rahat Iqbal is a Reader in Human-Centred Technology (HCT) and Director of Applied Computing Research Centre (ACRC) in the department of computing at Coventry University.

His research interests lie in information retrieval and supportive systems, in particular with regard to human factors and personalization of these systems based on user needs and context. Currently, he is the principal investigator of two funded projects including the EPSRC funded project tilted task sensitive search recommender system based on implicit feedback using intelligent techniques and a Technology Strategy Board funded project on driver's behavior with Jaguar Land Rover Ltd. He has published more than 70 papers in peer reviewed journals and reputable conferences and workshops. Dr. Iqbal is on the programme committee of several international conferences and workshops and organises the annual international workshop on Ubiquitous and Collaborative computing in conjunction with the BCS Conference on Human Computer Interaction. He is also a panelist and fellow of the UK Higher Education Academy and a member of the Nuffield Foundation (Science Bursaries & Oliver Bird Rheumatism Programme). Dr. Iqbal has also been editing several special issues of international journals within the field of information retrieval and user supportive systems.