3rd International Conference on Higher Education Advances, HEAd'17
Universitat Politècnica de València, València, 2017
DOI: http://dx.doi.org/10.4995/HEAd17.2017.5528

An Investigation into Third Level Module Similarities and Link Analysis

Keane, Michael[a]; Hofmann, Markus[b]
[a] Department of the Registrar, Institute of Technology Blanchardstown, Ireland;
[b]Department of Informatics, Institute of Technology Blanchardstown, Ireland.

Abstract

The focus of this paper is on the extraction of knowledge from data contained within the content of web pages in relation to module descriptors as published on http://courses.itb.ie delivered within the School of Business in the Institute of Technology Blanchardstown. We show an automated similarity analysis highlighting visual exploration options. Resulting from this analysis are three issues of note. Firstly, modules although coded as being different and unique to their particular programme of study indicated substantial similarity. Secondly, substantial content overlap with a lack of clear differentiation between sequential modules was identified. Thirdly, the document similarity statistics point to the existence of modules having very high similarity scores delivered across different years across different National Framework of Qualification (NFQ) levels of different programmes. These issues can be raised within the management structure of the School of Business and disseminated to the relevant programme boards for further consideration and action. Working within a climate of constrained resources with limited numbers of academic staff and lecture theatres the potential savings outside of the obvious quality assurance benefits illustrate a practical application of how text mining can be used to elicit new knowledge and provide business intelligence to support the quality assurance and decision making process within a higher educational environment.

Keywords: *web content mining, document similarity, text visualisation, network and link analysis.*

1. Introduction

One of the biggest challenges facing higher educational institutions is the exponential growth of educational data and the utilization of same to provide business intelligence to support the decision making process (Bala and Ojha, 2012). Educational data mining (EDM) is the application of data mining techniques to educational datasets and from this perspective the goal is not only to extract knowledge from data but also to use knowledge gained to improve the learning experience (Romero and Ventura, 2007; Romero and Ventura, 2010). Within the research field of EDM text mining has been employed for the purpose of analyzing the textual content of documents, forums, discussion boards and many others including web pages (Tane et al., 2004) the focus of this paper.

Web mining can generally be categorised into three areas (Zhang and Segall, 2008), web content mining, the aim of which is to extract knowledge from data contained within the content of web pages, web usage mining, the aim of which is to discover user access patterns from web usage logs and web link mining, the aim of which is to discover knowledge through the analysis of hyperlinks. The focus of this paper is on web content mining where an existing catalogue of module descriptors is extracted and then subsequently analysed using algorithmic techniques. Information in relation to programmes of study offered by the Higher Educational Institution under review is published in the public domain as per principles of the European Standards and Guidelines (ENQA, 2015) and available through http://courses.itb.ie. Module descriptors provide information on the educational aims/objectives, NFQ level and ECTS credits, module learning outcomes, the indicative content and assessment of modules, required reading, etc.

For the purpose of this analysis, the educational aims/objectives and the learning outcomes of modules within the School of Business are extracted and analysed using text mining techniques. The data mining objective is to acquire and extract the relevant information from the web page of each module, pre-process the data, perform document similarity analysis and export same for further analysis and text visualisation. The business objective is to quality assure module information as published within the public domain, identify degrees of commonality and overlap thereby identifying issues that may need to be addressed or provide further opportunity for the establishment of common modules that may be offered across the various disciplines within the school. Presently, programmes of study offered include General Business, Accounting and Finance, Sports Management, Business and Information Technology and International Business.

2. Data Aquisition

Rapidminer (Ritthoff et al., 2001) was employed for data acquisition, retrieving 231 web pages containing module information for modules delivered across the programmes within the School of Business. XPath was employed to extract the relevant information including the module code and title, aims/objectives and learning outcomes for each module. The bag of words approach was employed to model the extracted text. Generated tokens were transformed to lowercase and stop words were removed. The use of custom stopword lists that are domain specific, manually defined and that can be maintained by subject matter experts are considered good practice (Aggarwal et al., 2012) hence a custom stop word list was also applied to remove common words such as module, learner, student, completion, etc.

Examples of related approaches for modelling text document similarity include word based, keyword based and n-gram measures (Salton, 1989; Damashek, 1995). Having represented the extracted textual documents as term vectors the similarity between the documents was measured as the cosine of the angle between the vectors otherwise known as cosine similarity which was employed in generating document similarity statistics and is recognized as one of the most frequently used similarity measures employed for textual documents (Huang, 2008).

3. Document Similarity

Based on the document similarity statistics three issues of note became apparent. Firstly, modules although coded as being different and unique to their particular programme of study indicated substantial similarity. A total of 58 document pairings with a similarity score in excess of 89% point to the existence of common modules delivered across programmes within the School of Business. Examples include the Electronic Commerce modules delivered across the second year of the general business (BSST H2020) and the international business (INTB H2023) programmes returning a similarity score of 98% while the Supply Chain Management modules delivered across the fourth year of general business (BSST H4025) and accounting and finance programme (ACFN H4022) returned a similarity score of 99%.

The similarity of the Supply Chain Management modules are visually represented as word clouds in Figure 1. From a resourcing viewpoint one would expect that modules such as these with such a high similarity are recoded, retitled and offered as common within the same semester across the different programmes.

Secondly, there appears to be substantial content overlap with a lack of clear differentiation between sequential modules within each of the business programmes with 17 document

pairings of sequential modules showing document similarity scores in excess of 78%. Examples include the Sports Marketing modules 1 and 2 (SMCO H2014 and SMCO H2024) with a similarity score of 80% while the French Language Culture and Society modules 4a and 4b (INTB H4032 and INTB H4035) returned a similarity score of 88%.

The similarity of the Exercise Physiology 1 and 2 modules within the Sports Management programme are visually represented as word clouds in Figure 2.

A visual representation of the words unique to each of the Exercise Physiology modules is presented in Figure 3.

Figure 1: Word clouds of BSST H4025 (a) and ACFN H4022 (b)

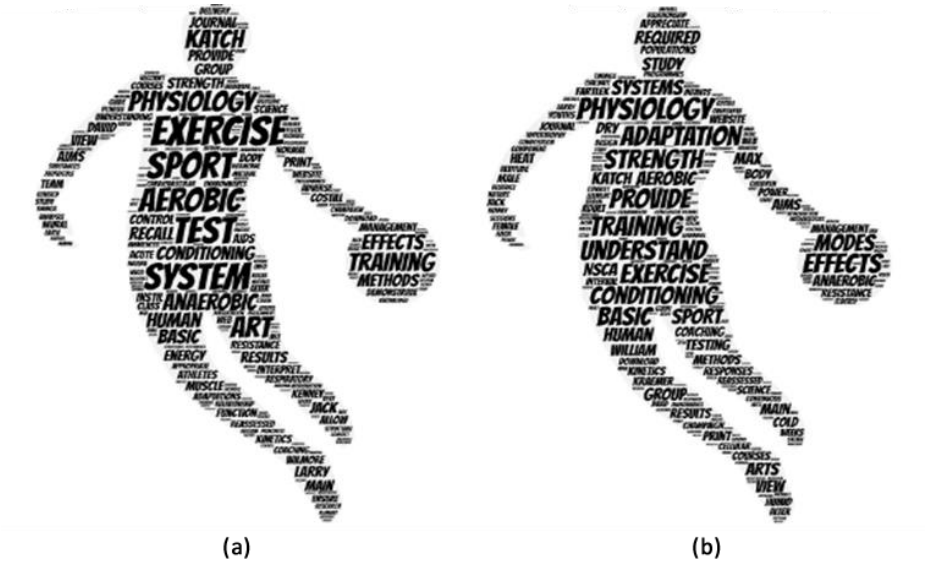

Figure 2: Word cloud of sequential modules Exercise Physiology 1 (a) and Exercise Physiology 2 (b) with a similarity score of 78%

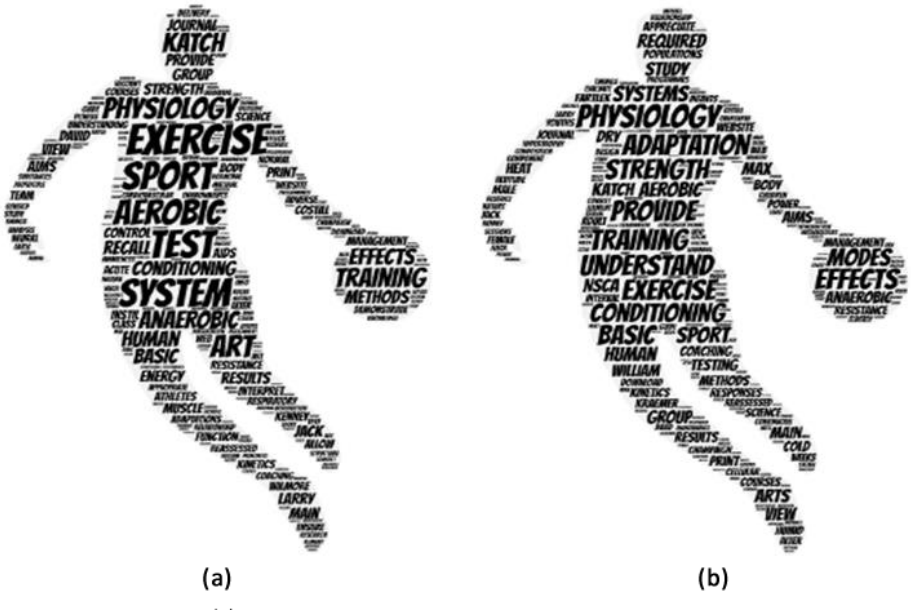

Figure 3: Word cloud of words unique to Exercise Physiology 1 (a) and Exercise Physiology 2 (b)

Thirdly and most concerning from a quality assurance perspective, the document similarity statistics point to the existence of modules having very high similarity scores delivered across different years, across different NFQ levels of different business programme disciplines. Examples of such programmes and modules are listed in Table 1.

Table 1. Modules with very high similarity scores delivered across different years, across different NFQ levels of different business programme disciplines.

Module 1	Module 2	Similarity
General Business - 4th Year BSST H4027 - Auditing	Accounting & Finance - 3rd Year ACFN H3013 - Auditing 1	97%
Sports Management - 3rd Year SMCO H3016 - Selling and Sales Management	International Business - 4th Year INTB H4030 - International Selling and Sales	96%

This third issue points to a possible failure in the quality assurance process in relation to module approval in that the learning outcomes of the modules in question do not accurately reflect the relevant award standard and NFQ level.

Module similarity as identified through similarity scores and visual techniques including word clouds can be further explored through network and link analysis and is discussed in the next section.

4. Link Analysis

In order to further analyse module similarity as provided by both the tabular output of scores and visualisation techniques discussed thus far, network and link analysis was employed using the interactive visualisation and exploration tool Gephi[♦]. While the document similarity statistics previously discussed provided the similarity between one module and another, Gephi was employed to identify the existence of multiple triangular relationships providing further evidence of the existence of previously undefined common modules being delivered within different programmes across the various disciplines within the School of Business.

[♦] https://gephi.org/

Examples include the Project Management module (Triangle B in Figure 3) delivered across the fourth year of the General Business (BSST H4014), International Business (INTB H4012) and the Business and Information Technology (BSIT H4013) programmes. The Introduction to Management and Cost Accounting module (Triangle A in Figure 3) delivered across the second year of the General Business (BSST H2022), Accounting & Finance (ACFN H2012) and the Business and Information Technology (BSIT H2017) programmes. Also, the Electronic Commerce module (Triangle C in Figure 3) delivered across the second year of the General Business (BSST H2020), International Business (ACFN H2023) and the Sports Management and Coaching (SMCO H2016) programmes.

Visual representation of these examples of triangular relationships with similarity scores using Gephi is presented in Figure 3 on the following page. Expanding the data capture to include modules from other faculties/schools could possibly highlight further multidimensional relationships between programmes allowing further rationalisation and more efficient use of resources.

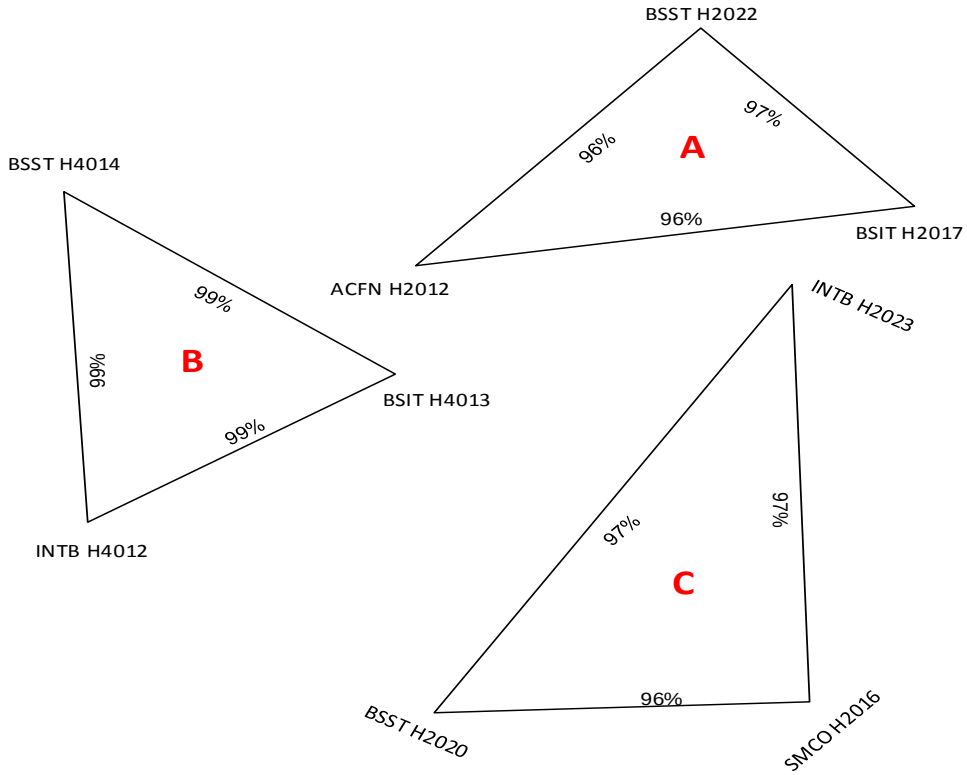

Figure 3: Triangular relationships with similarity scores using Gephi

5. Conclusion

Having gathered the necessary module data, extracted the relevant information and pre-processed the data, document similarity statistics were generated and further analysed through various visual exploration techniques. Resulting from this analysis, three issues of note became apparent. Firstly, modules although coded as being different and unique to their particular programme of study indicated substantial similarity. Based on this analysis 58 document pairings with a similarity score in excess of 89% point to the existence of common modules delivered across programmes within the School of Business. Secondly, substantial content overlap with a lack of clear differentiation between sequential modules was identified through this analysis with 17 document pairings having a similarity score in excess of 78%. Thirdly, the document similarity statistics point to the existence of modules having very high similarity scores delivered across different years across different NFQ levels of different programmes with 4 document pairings identified. These issues can now be raised within the management structure of the School of Business and disseminated to the relevant programme boards for further consideration and action. Working within a climate of constrained resources with limited numbers of academic staff and lecture theatres, the potential savings outside of the obvious quality assurance benefits illustrate a practical application of how text mining can be used to elicit new knowledge and provide business intelligence to support the quality assurance and decision making process within a higher educational environment.

References

Aggarwal, C.C. and Zhai, C., 2012. *Mining text data*. Springer Science & Business Media.

Bala, M. and Ojha, D.B., 2012. Study of applications of data mining techniques in education. *International Journal of Research in Science and Technology*, *1*(4), pp.1-10.

Damashek, M., 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, *267*(5199), p.843.

ENQA 2015. *Standards and guidelines for quality assurance in the European Higher Education Area (ESG)*. Brussels, Belgium.

Huang, A., 2008, April. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (pp. 49-56).

Ritthoff, O., Klinkenberg, R., Fischer, S., Mierswa, I. and Felske, S., 2001, October. Yale: Yet another learning environment. In *LLWA 01-Tagungsband der GI-Workshop-Woche, Dortmund, Germany* (pp. 84-92).

Romero, C. and Ventura, S., 2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), pp.135-146.

Romero, C. and Ventura, S., 2010. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), pp.601-618.

Salton, G., 1989. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*.

Tane, J., Schmitz, C. and Stumme, G., 2004, May. Semantic resource management for the web: an e-learning application. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 1-10). ACM.

Zhang, Q. and Segall, R.S., 2008. Web mining: a survey of current research, techniques, and software. International Journal of Information Technology & Decision Making, 7(04), pp.683-720.