

Applying Text Mining to Analyze the Internationalization Levels of Universities in Taiwan

Hongfeng Lai, Siouying Chen, Shuhuei Huang, Hueimei Lin, Yunwei Guo

Dept. of Business Management, National United University, Miaoli city, Taiwan, China.
Email: hflai@nuu.edu.tw

Received June, 2013

Abstract: In the era of globalization, the university managers face more challenges such as international competition, rapidly changing education market. Internationalization of education has become an important strategy for education managers to adopt. Internetization is necessary conditions for Universities to carry out internationalization. The campus information portal, which supports an exchanging platform for other country's potential students and scholars through information transparency, has become observation focus of internationalization of university. The completeness of campus information portal is a necessary condition for internationalization. In this paper, we focus on general university in Taiwan, screening with five or more college. The source documentaries text are collected from the campus information portal in English version, and apply text mining techniques to analyze the internationalization levels and clusters of universities in Taiwan. The results might help university administrators to improve or enhance the internationalization levels, and provide reference for mutual learning between universities.

Keywords: Text Mining; Internationalization; Internetization

利用文件探勘技术分析台湾普通大学的国际化层次

赖宏峯, 陈秀英, 黄淑惠, 林蕙玫, 郭竣玮

联合大学经营管理学系, 苗栗市, 台湾, 中国
Email: hflai@nuu.edu.tw

收稿日期: 2013年6月

摘要: 在全球化的时代, 国际竞争益形激烈, 教育市场快速变迁, 在大学经营上面临更多挑战, 教育国际化成为教育管理者采取的重要策略之一。大学推行国际化, 网络化为其必要条件, 经由校园信息入口网站透明化校园信息, 提供本国与他国潜在学生与学者交流的平台, 使校园信息入口网站成为大学推行国际化成效的观察重点, 良备的大学校园信息入口网站为国际化的必要条件。本文以台湾普通大学为案例, 筛选具五间学院以上之普通大学校园信息入口网站之英文网页为文件来源, 利用文件探勘技术分析台湾普通大学的国际化层次, 与推行国际化的相似性; 本研究成果可供校务管理者改善或提升国际化层次之依据, 亦可作为普通大学间相互学习之参考。

关键词: 文件探勘; 国际化; 网络化

1. 前言

在全球化的时代, 国际竞争激烈、教育快速变化、少子化, 使大学经营面对更多挑战, 教育国际化成为

教育革新的重要课题与趋势^[1,2]。因此, 各大学经由发展国际化之教育能力、提升办学素质与改善竞争力, 以吸引全球潜在的毕业生。台湾近年来积极鼓励大学

提升国际化,且将国际化程度列为大学校务评鉴项目之一,因此各大学投入相当比例经费于发展国际化,政大未来交易所最近公布台湾各大学之相关排名^[3],其国际化被列为交易项目之一,可见国际化之议题为目前所关注的焦点,然而未来交易所中大学国际化的排名是按网友依据印象下单给分,并没有客观的依据,如何建立衡量台湾大学国际化的客观标准,值得进一步探讨。

高等教育负有培育高水平且优质人力的使命与功能,是推动国家永续发展与提升国际竞争力和经济实力的主因之一,直接影响国家社会之整体发展^[4]。

网络化可做为新兴教育国际化的必要条件^[5],大学透过校园信息入口网站作为国际化之平台,推动国际化呈现不同的结果;由于网络数据本质上量多且杂,要从中找出可用信息,须费相当作业成本,为克服此困难,本文藉由文件探勘软件,从大量文件中,调查分析台湾普通大学推行国际化的层次与结果。

2. 文献回顾

台湾面临瞬息万变的国际外在环境,以及人才流动无国界的趋势,为教育带来的冲击主要有三:在知识经济时代中,面临终身学习之需求;在数位信息时代中,造成学习型态的转变;在全球化时代中,国际竞争益形激烈。

教育产业日益重视国际化课题,藉此招收外国学生与我国学生交流学习,以及培养学生具备国际人才的实力,推动教育国际化已成为必要的发展趋势。

在网络化的过程中,透过因特网技术可传播、应用信息,藉由信息入口网站技术,以交流互动方式将组织推向国际。国际化的组织必然建置网络化的环境,但已建置网络化环境的组织未必达成国际化目标,因为要达成国际化所需的条件更严格,包括多元之文化互动与交流^[5]。

目前台湾各大学所做的国际化相关活动内容,一般会在其校园信息入口网站公布,所以透过校园信息入口网站之文件探勘,可得知当前各大学国际化层次。

教育国际化是国家发展国际化重要指标之一,依据《大学教育政策白皮书草案》指出台湾高等教育应

积极参与国际组织,因应国际市场的竞争,以提升国际竞争力^[6],以竞争力为焦点,高等教育不仅关注台湾国际学者与学生的比例或是人才交换,进一步更至其他国家开设课程、设立双联学制,积极拓展海外市场以招募人才。

3. 研究方法

3.1. 文件探勘技术

文件探勘为一种分析大量文件的技术,可提供使用者某些特定的信息,发现某些特征及其间的关联性。文件探勘主要藉由传统信息检索、机器学习与人工智能等技术,企图在大量的文件中归纳及推衍隐含的信息与知识^[7]。

一般文件探勘的程序包括七大流程:文本选择与处理、断词与长词修正、词频计算、权重计算、群集分析(分类分析、关系型法则分析)、成果诠释、与成果图像化等。

文本选择与处理指将所要处理的文件进行分割整理,移除非研究标的,可大幅减少系统工作量。

断词与长词修正指将切割完成后之进行断词处理,并对断词后的结果进行检查。

词频计算指针对每一个文本断词后的词汇,将非名词及动词的词汇删除后,计算各词汇出现的频率,在每一个文本都经过相同的处理后,以 m 文件数 \times n 词数的矩阵表示。

权重分析指是常用于信息检索与文本挖掘的加权技术,权重分析的结果以 $m \times n$ 矩阵表格呈现供分类程序进行计算^[8],常用的指标有词频与逆向文件频率。词频(Term Frequency: TF)指的是特定的词语在该文件中出现的次数,如(1)式所示,其中 $n_{i,j}$ 是该词在文件 d_j 中的出现次数,而分母则是在文件 d_j 中所有字词的的出现次数之和。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad (1)$$

逆向文件频率(Inverse Document Frequency: IDF)测量词语普遍重要性如式(2),其中 $|D|$ 表示文件总数, $|\{j:t_i \in d_j\}|$ 表示包含词语 t_i 的文件数目。

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \quad (2)$$

群集分析 (Clustering analysis) 指将权重计算的矩阵表透过类神经网络方法 GHSOM 或 K-mean 统计方法, 分析所产生之群集结果^[9]。

成果图像化指将最后分类的成果以图的形式呈现。

成果诠释指针对分类的结果进行深入的质化分析与探讨, 其探讨的模式主要有以下两个主轴: 分析自动分类结果之合理性, 与列出各分类的重要关键字。

3.2. 文件探勘应用工具—QDA Miner

QDA Miner 是 Provalis research 公司所开发之文件探勘应用软件, 整合 Word Stat 和 Sim Stat 等相关探勘软件, 提供统计、群集、序列分析等工具, 能快速分析与探索文件背后信息, 是本文选择 QDA Miner 的主要原因。

QDA Miner 可从图形、卷标、检索等部分审查每篇文件的相似性, 也可把关键词归类为不同领域 (category), 并分析文件是否有在此领域中。QDA Miner 把所有文件当成项目, 从这些项目中将每篇文件断词, 再以断词进行出现次数等统计、分析并计算出其 TF-IDF 等信息。QDA Miner 提供广泛的探索工具, 适用于多种文件格式, 本文则是采用 Word 文件格式。

依照文件所选择的变量作为搜寻依据以分析档案, 再针对代码分析, 获得搜寻结果。此分析方法可由 QDA 软件的分析 (Analyze) 选项里选择代码频率 (Coding frequency)。

如表 1 为代码频率范例表, 其中范畴 (Category) 指代码所属之集合, 如 Partner Universities 是属于 Category: 1。计数 (Count) 指某一代码 (Code) 在所有档案中出现次数, 例如 Partner Universities 总出现次数为 47。代码比例 (%Codes) 指 $\text{Count} / \sum \text{count} \times 100\%$, 案例数 (Cases) 指包含某代码的案例数目。案例数比例 (%Cases) 指 $\text{Cases} / \text{总文章数} \times 100\%$ 。

Table 1. Example of code frequency analysis
表 1. 代码频率范例表

Category	Code	Description	Count	% Codes	Cases	% Cases
1	Partner Universities		47	7.5%	37	100.0%
1	English learning environment		56	9.0%	37	100.0%
1	Chinese Language Course		37	5.9%	30	81.1%
2	summer program		33	5.3%	23	62.2%
2	International Cooperation		43	6.9%	37	100.0%

4. 研究设计

4.1. 案例介绍

普通大学(综合大学)是高等教育重要的一环, 普通大学是以学术导向为目标, 衔接各阶段之普通教育, 从重视每个学科均衡发展之普通高中教育, 到重视理论研究与应用科学之普通大学教育, 建构成一个完整的普通教育体系, 本文是以普通大学为研究对象。

本文将以 37 所台湾普通大学校园信息入口网站英文网站之文件数据, 运用文件探勘软件(QDA Miner)分析, 依照本文所订之国际化层次架构, 从样本普通大学所提供之文件, 将此分类结果提供给校务管理者了解当前国际化层次, 亦可作为样本普通大学间相互学习之参考。

4.2. 样本选择

本文选择样本普通大学之依据有二, 其一为所选非特定宗教学校, 其二为所选需具五间学院以上相当规模的普通大学, 结果共有 37 所, 做为比较台湾普通大学国际化层次的数据源。

4.3. 国际化层次定义

本文依据教育部补助办理国际学术教育交流活动实施要点, 依照大学投入国际化之资源多寡, 并参考程荣凯(2008)所提出的科技大学国际化层次五阶段^[10], 建立本文国际化层次的区分模式。

第一层指启动阶段层次: 指台湾普通大学推动与国外大学签订姐妹校、制定外语学习环境(提供外文网页、外语相关检定等)、提供以英文教学之华语课程。

第二层指各类合作与交流层次: 指台湾普通大学推动校内学生出国研习、游学(Study Abroad)、举办台湾外夏令营等活动(summer program)、提供办理国际合作与交流事务(如与国外学校签订合作协议书)、交换学生(Income or Outcome)或攻读学士、硕士与博士学位(International Degree Student)、推动学生赴国外工读、实习、志工等活动。

第三层指课程、教学与研究国际化层次: 指台湾普通大学推动校内师生出席国际会议、教师开授全英语授课的专业课程、办理境外教学学分班、办理国际

学术研讨会、国外访问学者驻校研究、交换教授、教师担任国际著名期刊编辑委员等项目。

第四层指校园国际化层次：指台湾普通大学签订与实施双联学位、拟定近程与远程目标积极推动校园国际化(英文导览 Living、Campus Life)、将国际事务提升为一级单位、延聘长期专业领域外籍教师等。

第五层指商业化层次：指台湾普通大学建立国外分校、建立网络虚拟大学、办理大学国际化营销相关活动(如举办国际教育展、大学博览会等)、和国外大学策略联盟(如讲座、博士后研究等)项目。

5. 文件探勘结果与发现

依照代码频率(Coding Frequency)分析,从代码(Code)方面讨论,在全部的层次内制定英语学习环境(English learning environment)占 8.8%, 是被提及最多的代码, 并且与姊妹校(Partner School)提供办理国际合作与交流事务(International Cooperation)在所有的案例中都有提到, 表示 37 所大学都很重视这三项国际化活动。

相较之下在境外教学学分班(Overseas teaching course)、教授担任国际知名期刊编辑委员(International journal editor)在所有案例中都没有提到, 其中网络虚拟大学(Network Virtual University)因为达成条件较为严苛所以各大学目前都没有做到。

结果显示一般教授担任国际知名期刊编辑委员(International Periodical)与境外教学学分班(Overseas teaching course)不容易在网站上发现其相关内容, 所以建议各大学可以在网站上公布相关信息以方便国际学生可以了解学校国际化相关活动, 如图 1 所示。

Category	Code	Description	Count	% Codes	Cases	% Cases
1	Chinese Language Course		38	6.1%	30	81.10%
1	English learning environment		55	8.80%	37	100.00%
1	Partner Universities		47	7.50%	37	100.00%
1 total			140	22.40%		
2	Exchange Student		45	7.20%	34	91.90%
2	International Ambassador		28	4.50%	20	54.10%
2	International Cooperation		43	6.90%	37	100.00%
2	International Degree Student		33	5.30%	29	78.40%
2	Study Abroad		28	4.50%	26	70.30%
2	summer program		33	5.30%	23	62.20%
2 total			210	33.70%		
3	English-taught		27	4.30%	22	59.50%
3	Faculty Exchange		30	4.80%	22	59.50%
3	International Periodical		0	0%	0	0%
3	Overseas teaching courses		0	0%	0	0%
3	Symposium		25	4.00%	22	59.50%
3	Visiting Scholar		27	4.30%	22	59.50%
3	international conferences		34	5.40%	26	70.30%
3 total			143	22.80%		
4	Campus Life		32	5.10%	31	83.80%
4	Dual degree programs		25	4.00%	23	62.20%
4	Full-time Distinguished Professor		6	1.00%	6	16.20%
4	Office of International Affairs		33	5.30%	31	83.80%
4 total			96	15.40%		
5	Education Fair		10	1.60%	9	24.30%
5	Lecture		14	2.20%	14	37.80%
5	Network Virtual University		0	0%	0	0%
5	Postdoctoral Research		10	1.60%	10	27.00%
5	overseas campus		1	0.20%	1	2.70%
5 total			35	5.60%		

Figure 1. Results of Code Frequency Analysis
图 1. 代码频率分析结果

依照代码频率(Coding Frequency)分析如表 2, 从国际化层次(Category)角度, 在第一层: 启动阶段占 22.40%、第二层: 各类合作与交流占 33.70%、第三层: 课程教学研究国际化占 22.80%, 多数集中在这三层上, 显示台湾各大学以将前三层活动列为国际化基本条件。

第 5 层商业化层次只占 5.60%, 在所有的层次中属于少数, 因为要达到商业化层次的条件所耗费资源较为巨大, 所以各大学比较难达到, 建议各大学可以加强商业化层次以加强国际化深度。

6. 结论

本研究所建构的国际化层次分类上, 初步获得不错的效果, 但仍有其限制, 尤其在属于国际化活动内含于其他活动, 如论文成果来自跨国合着, 此类文件并无法在大学入口网站上取得, 因此自动分群技术并无法具体有效呈现这类文件的结果。

本文搜集各普通大学入口网站之数据, 运用文件探勘技术分析其网页所显示之国际化相关信息, 经由本文所提出之国际化层次分类, 调查与分析台湾普通大学院校目前的国际化层次, 提供校务管理者发掘文件背后所隐藏的信息。透过文件探勘软件建构出透明化信息, 可帮助校务管理者了解当前其国际化层次, 作为改善或提升国际化之依据。

Table 2. Summing-up of code frequency analysis
表 2. 统整后的代码频率分析结果

Category	Code	Description	Count	%Codes	Cases	%Cases
1	Chinese Language Course		38	6.10%	30	81.10%
1	English learning environment		55	8.80%	37	100.00%
1	Partner Universities		47	7.50%	37	100.00%
1 total			140	22.40%		
2	Exchange Student		45	7.20%	34	91.90%
2	International Ambassador		28	4.50%	20	54.10%
2	International Cooperation		43	6.90%	37	100.00%
2	International Degree Student		33	5.30%	29	78.40%
2	Study Abroad		28	4.50%	26	70.30%
2	summer program		33	5.30%	23	62.20%
2 total			210	33.70%		
3	English-taught		27	4.30%	22	59.50%
3	Faculty Exchange		30	4.80%	22	59.50%
3	International Periodical		0	0%	0	0%
3	Overseas teaching courses		0	0%	0	0%
3	Symposium		25	4.00%	22	59.50%
3	Visiting Scholar		27	4.30%	22	59.50%
3	international conferences		34	5.40%	26	70.30%
3 total			143	22.80%		
4	Campus Life		32	5.10%	31	83.80%
4	Dual degree programs		25	4.00%	23	62.20%
4	Full-time Distinguished Professor		6	1.00%	6	16.20%
4	Office of International Affairs		33	5.30%	31	83.80%
4 total			96	15.40%		
5	Education Fair		10	1.60%	9	24.30%
5	Lecture		14	2.20%	14	37.80%
5	Network Virtual University		0	0%	0	0%
5	Postdoctoral Research		10	1.60%	10	27.00%
5	overseas campus		1	0.20%	1	2.70%
5 total			35	5.60%		

本研究之结果可提供给大学校务管理者了解当前国际化层次，可作为普通大学间相互学习之参考。

参考文献 (References)

- [1] 张钦盛. 台湾教育国际化的政策与实施[J]. 教育资料与研究双月刊, 2006, 71: 1-16.
- [2] J. Lee and M. You. Does internationalization have its own face in each country? measuring the internationalization of web presence on higher education admission web pages between USA and Taiwan. Internationalization, Design and Global Development, P. L. P. Rau, ed., Springer, 2011: 524-532.
- [3] 政大未来交易所.
http://www.nownews.com/2013/05/22/91-2941903_1.htm, 2013 5-20.
- [4] 叶至诚. 高等教育发展的策略与愿景[J]. 台北: 扬智文化, 2002.
- [5] H. Etemad, I. Wilkinson and L. Dana. Internetization as the necessary condition for internationalization in the newly emerging economy. Journal of International Entrepreneurship, 2010, 8 (4): 319-342. [doi:10.1007/s10843-010-0062-x](https://doi.org/10.1007/s10843-010-0062-x)
- [6] 杨巧玲. 高等教育国际化的意义与争议. 教育政策论坛, 2004, 7 (1): 101-110.
- [7] D. Sullivan. Document warehousing and text mining: techniques for improving business operations, marketing, and sales, John Wiley & Sons, Inc., 2001.
- [8] L. Jing, H. Huang and H. Shi. Improved feature selection approach tfidf in text mining. The First International Conference on Machine Learning and Cybernetics, 2002: 944-946. [doi:10.1109/ICMLC.2002.1174522](https://doi.org/10.1109/ICMLC.2002.1174522)
- [9] 陈文华, 施人英, 吴寿山. 探讨文字采掘技术在管理者知识地图之应用. 中山管理评论, 2004: 35-64.
- [10] 程荣凯. 科技大学教育国际化指标建构之研究[D]. 台北市: 台湾师范大学工业教育学系学位论文, 2008.