Leveraging Scholarly APIs for Bulk Citation Analysis of Collection Availability

PRESENTER: Aaron Tay
Library Analytics Manager
Singapore Management University

INTRO:

How can we modernize use of citation analysis in collection evaluation? Instead of manual samples can we automate the whole process with APIs? Given that 28%-42% of articles are now Open Access (Piwowar et. al, 2018), how can we that that into account? Can we use the latest machine learning e.g. Clustering techniques to make sense of what is cited?

METHODS

1. Used Scopus to extract all citations made from SMU papers in 2017 & 2018
2. Used APIs from Primo, Unpaywall, Open Access button , Google books to check availability.
3. A human checked 500 samples as a Gold Standard
4. Calculated % available, both free only & free + in library collection & adjust
5. Attempted to enhance abstracts and full text using CORE
6. Visualized data using VOSviewer (term map) and Topic modelling (LSA)

RESULTS

- While recall (86.7%) & precision (98.8%) for Primo library collection check is reasonable the false negative rate once you included free material was poor (65.6%)

Collection assessment using citation analysis automated by APIs checks have **acceptable accuracy** for **library collections only** (98.8% precision, 86.7% recall) but have **high false negatives** for **free to read items**.

	ALL API = Available	ALL API = Not available	Total	% Correct
Available (Gold Standard)	352	95	447	78.7%
Not Available (Gold Standard)	3	50	53	94.3%
Total	355	145	500	
	99.0%	34.4%		

Take a picture to **download** the **full presentation**

Sample 500 random sample check in Primo

Accuracy of Primo API vs Human check in Primo

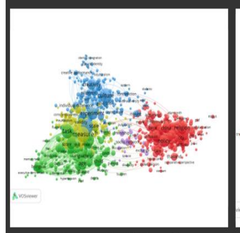
	Primo API = available	Primo API = Not available	Total	% Correct
Available in Primo (Gold standard)	327	50	377	86.7%
Not available in Primo (Gold Standard)	4	119	123	96.7%
Total	331	169	500	
% correct	98.8%	70.4%		

API	Citations worked on	Found	% Found	Comment	Adjusted %
Primo	34,617	22,060	63.7%	Everything	73.60%
UnPayWall	20,736	5,333	25.7%	DOI item only	33.90%
Google Books	636	111	17.5%	Book item only	-
OpenAccessButton	13,737	868	6.30%	Check on remaining unfound	-

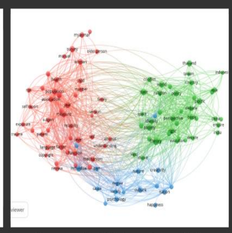
Preliminary results by School (all)

School	Reference Cited by School	Reference Cited by School with FullText Availability (Yes)	%available (including free)	% adjusted
LKCSB	9,245	7,492	81	92.7
SIS	12,248	7,055	57.6	84.9
SOA	1,546	1,268	82	93.1
SOE	2,837	2,229	78.6	91.9
SOL	1,772	698	39.3	78.7
SOSS	5,468	3,942	72	89.7
Total	33,116	22,684		

Visualization using Vosviewer

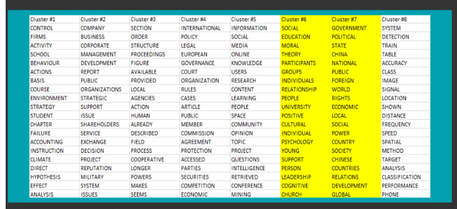


Term map - Social Science all



Term map - Social Science - Not available

Visualization using topic models (LSA)



Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5	Cluster #6	Cluster #7	Cluster #8
CONTROL	COMPANY	SECTION	INTERNATIONAL	INFORMATION	SOCIAL	GOVERNMENT	SYSTEM
FIRMS	BUSINESS	ORDER	POLICY	SOCIAL	EDUCATION	POLITICAL	DETECTION
ACTIVITY	CORPORATE	STRUCTURE	LEGAL	MEDIA	MORAL	STATE	TRAIN
SCHOOL	MANAGEMENT	PROCEEDINGS	EUROPEAN	ONLINE	THEORY	CHINA	TABLE
BEHAVIOUR	DEVELOPMENT	FIGURE	GOVERNANCE	KNOWLEDGE	PARTICIPANTS	NATIONAL	ACCURACY
ACTIONS	REPORT	AVAILABLE	COURT	USERS	GROUPS	PUBLIC	CLASS
BASIS	PUBLIC	PROVIDED	ORGANIZATION	RESEARCH	INDIVIDUALS	FOREIGN	IMAGE
COURSE	ORGANIZATIONS	LOCAL	RULES	CONTENT	RELATIONSHIP	WORLD	SIGNAL
ENVIRONMENT	STRATEGIC	AGENCIES	CASES	LEARNING	PEOPLE	RIGHTS	LOCATION
STRATEGY	SUPPORT	ACTION	ARTICLE	PEOPLE	UNIVERSITY	ECONOMIC	SHOWN
STUDENT	ISSUE	HUMAN	PUBLIC	SPACE	POSITIVE	LOCAL	DISTANCE
CHAPTER	SHAREHOLDERS	ALREADY	MEMBER	COMMUNITY	CULTURAL	SOCIAL	FREQUENCY
FAILURE	SERVICE	DESCRIBED	COMMISSION	OPINION	INDIVIDUAL	POWER	SPEED
ACCOUNTING	EXCHANGE	FIELD	AGREEMENT	TOPIC	PSYCHOLOGY	COUNTRY	SPATIAL
INSTRUCTION	DECISION	PROCESS	PROTECTION	PROJECT	YOUNG	SOCIETY	METHOD
CLIMATE	PROJECT	COOPERATIVE	ACCESSED	QUESTIONS	SUPPORT	CHINESE	TARGET
DIRECT	REPUTATION	LONGER	PARTIES	INTELLIGENCE	PERSON	COUNTRIES	ANALYSIS
HYPOTHESIS	MILITARY	POWERS	SECURITIES	RETRIEVED	LEADERSHIP	RELATIONS	CLASSIFICATION
EFFECT	SYSTEM	MAKES	COMPETITION	CONFERENCE	COGNITIVE	DEVELOPMENT	PERFORMANCE
ANALYSIS	ISSUES	SEEMS	ECONOMIC	MINING	CHURCH	GLOBAL	PHONE

Topic model using title/abstract from Scopus + CORE API

SMU SINGAPORE MANAGEMENT UNIVERSITY
Libraries