

Research and the Long Tail: A Large Scale Citation Analysis

**Wu, Ling-ling, Luesak
Luesukprasert, and Lynne Lee
National Taiwan University**

Citations with availability of online articles

- **Importance of citation:**
 - Estimation of research impacts
 - **Concentration of citation: (figure 1)**
 - 50% of citations comes from 15% of publications; 90% of citations comes from 50% of publications (Seglen, 1992)
 - **Availability of online articles:**
 - Lower cost of information seeking
 - Powerful facilitations (recommendation)
 - Different reading behavior (keyword search)
- ➔ **Question: how does availability of online articles affect concentration of citation?**

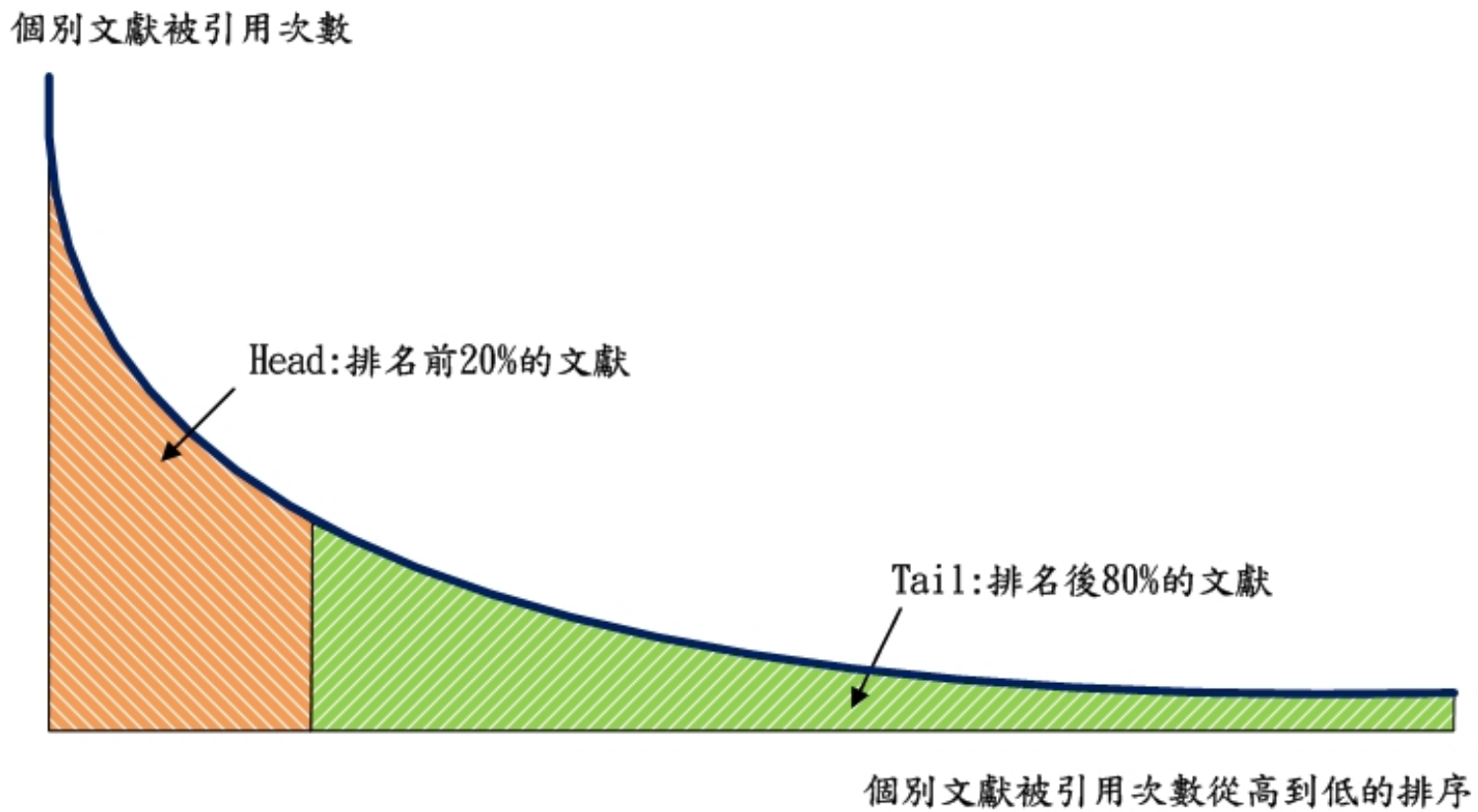


Figure 1. Frequency of citations ranked from high to low

Related theories

- **Long tail theory: (Anderson, 2004, 2006)**
 - The Internet helps us to find what we really need.
 - Little limits on display space

The market will be divided into many little niche markets
(Brynjolfsson, et al., 2007; Brynjolfsson, et. al., 2006)

➔ predicts less concentration, tail becomes longer and fatter after applications of the Internet
- **Winner-take-all theory: (Frank, 1996)**
 - The Internet helps us to find really high quality products.
 - Information becomes more transparent.

Everyone goes for the best. (Elberse & Oberholzer-Gee, 2006)

➔ predicts greater concentration, tail becomes shorter and thinner after applications of the Internet

Information seeking in Research

- **Fragmented interests**
- **Journals have their own specific focus even within the same field, e.g. marketing (Tellis, Chandy, & Ackerman, 1999)**
- ➔ **People go for the best quality paper, but related to their research interests.**

Hypotheses

- **Hypothesis 1: The proportion of tail articles to be cited increases and the proportion of head articles to be cited decreased after the introduction of online articles.**
- **Hypothesis 2: The growth rate of citations to tail articles exceeds that of citations to head articles after the introduction of online articles.**

Methodology

- **Material:**
 - **Citee:** articles searched by Google scholar in the field of information behavior (key word search)
 - **Citer:** articles that cite the citees described above
 - Within the field of information behavior
 - Beyond the field of information behavior
- **Citation window: 10 years time frame**
 - **Pre-Eresearch period: citee 1970-1980 vs. citer 1980-1990**
 - **Eresearch emergent period: citee 1980-1990 vs. citer 1990-2000**
 - **Eresearch Prevalent period: citee 1990-2000 vs. citer 2000-2008**

Head vs. tail citees

- **Cite articles (citer):** articles that cite other articles
- **Core articles (citee):** articles being cited by other articles
 - **Head articles:** top 20% of the articles that receive the most citations (popular, hit citees)
 - **Tail articles:** the rest articles
 - **Tail articles with impacts:** tail articles that receives at least one citation

- **Analysis: manifest content analysis**
 - **Proportion of citations accounted by head vs. tail citee articles**
 - ➔ **The larger proportion of citations accounted by head indicates greater concentration**

Table 1. Numbers of core articles and cite articles.

Data Set	Number of Core Articles		Number of Cite Articles that are Cite Articles (Citations by papers <u>within Information Behavior</u>)		Number of Cite Articles that are Not Core Articles (Citations by papers <u>beyond Information Behavior</u>)	
	Total	% of Total	Total	% of Total	Total Count	% of Total
1970s	478	2.83%	917	4.07%	5,010	5.38%
1980s	1,086	6.43%	4,030	17.88%	15,160	16.29%
1990s	4,051	23.99%	9,725	43.15%	39,467	42.40%
2000s	11,271	66.75%	7,864	34.90%	33,449	35.93%
Total	16,886	100.00%	22,536	100%	93,086	100%

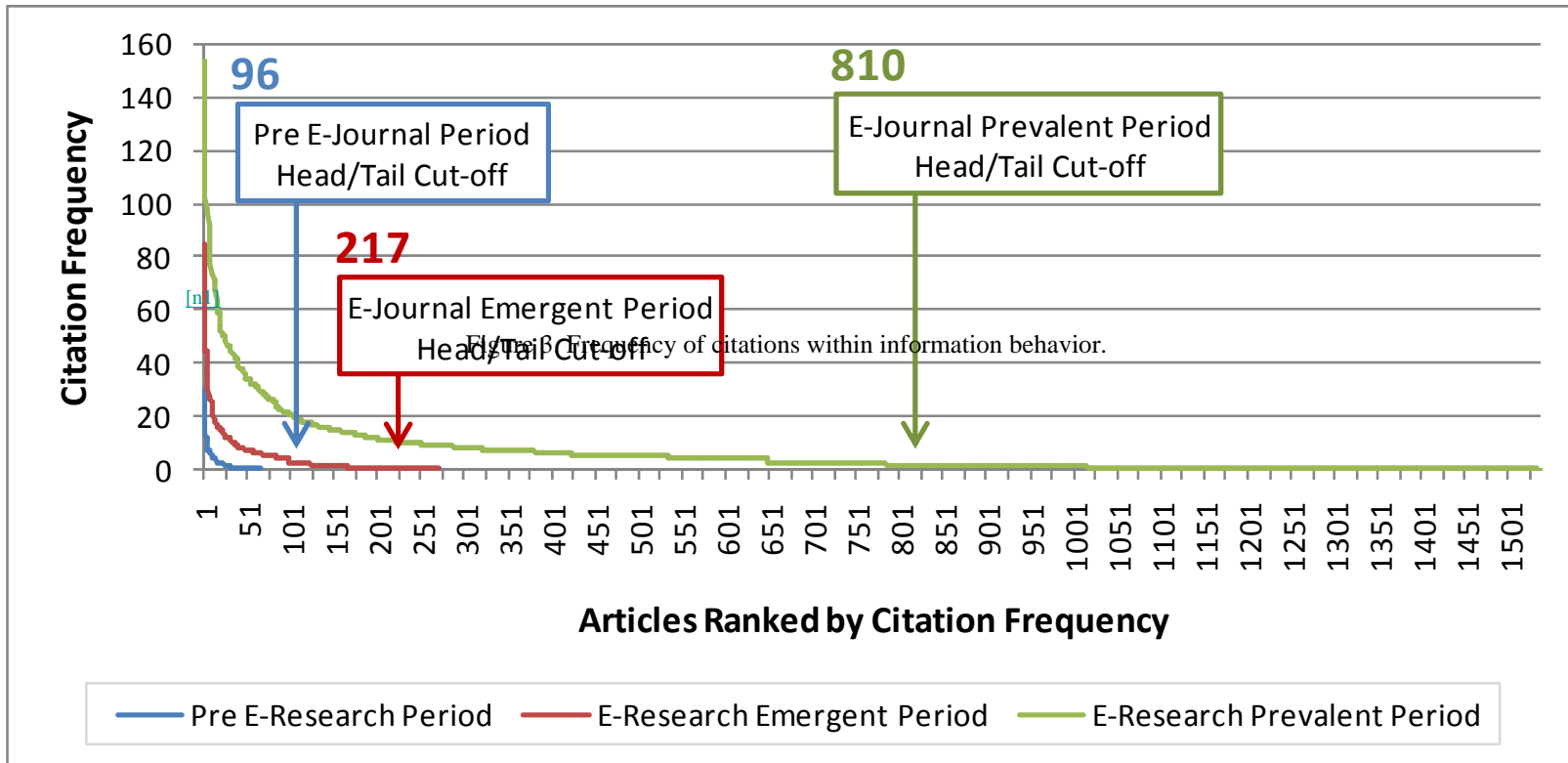


Figure 1. Frequency of citations within information behavior.

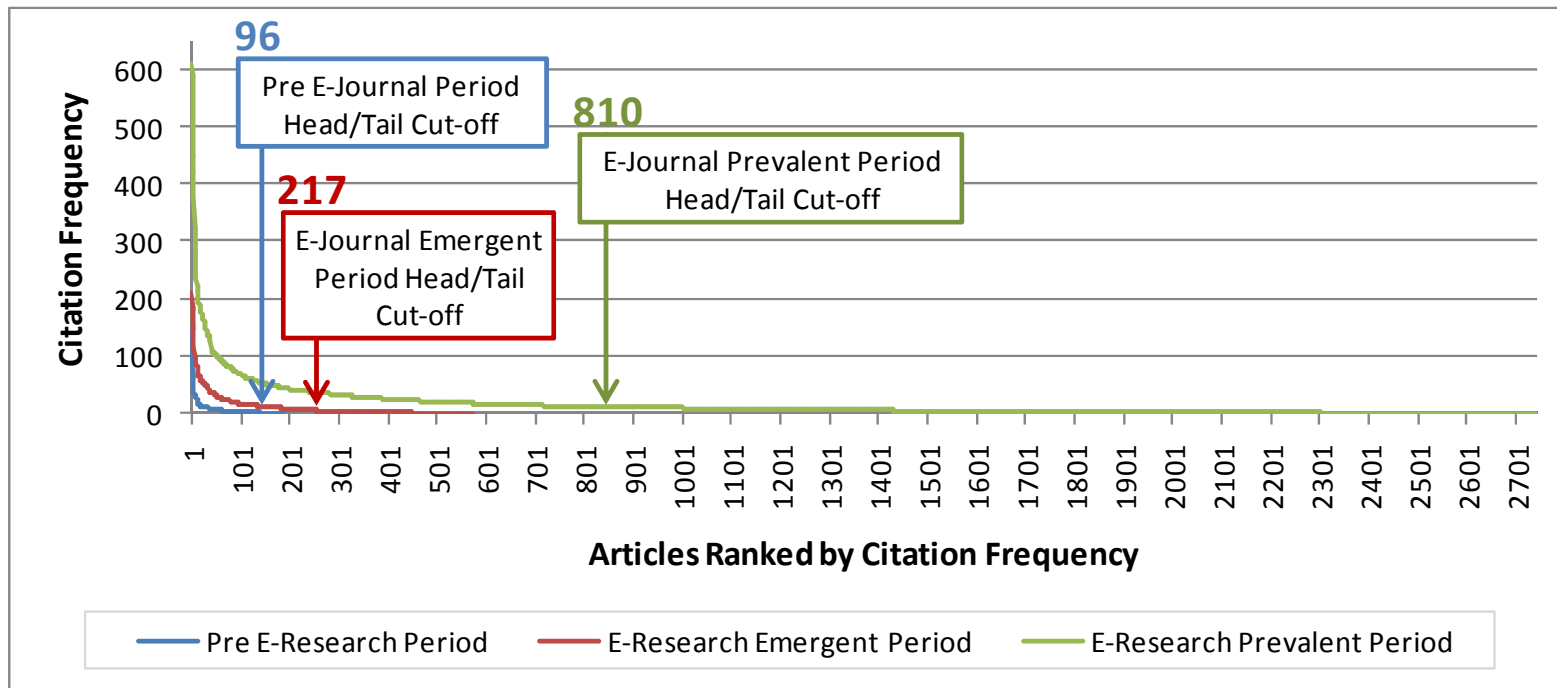


Figure 2. Frequency of citations beyond information behavior.

Table 2. Citations to head and tail articles within information behavior.

Periods		Total	Head Articles	Tail Articles	Tail Articles with Impact
Pre E-Research	Core Articles	478	20% (96/478)	80% (382/478)	0.00% (0/382)
	Number of Citations	218	100% (218/218)	0.00% (0/218)	
	Average Number of Citations	0.46	2.27	0.00	
E-Research Emergent	Core Articles	1,086	20% (217/1,086)	80% (869/1,086)	6.10% (53/869)
	Number of Citations	1,400	96.21% (1,347/1,400)	3.79% (53/1,400)	
	Average Number of Citations	1.29	6.21	0.06	
E-Research Prevalent	Core Articles	4,051	20% (810/4,051)	80% (3,241/4,051)	22.31% (723/3,241)
	Number of Citations	10,050	90.76% (9,121/10,050)	9.24% (929/10,050)	
	Average Number of Citations	2.48	11.26	0.29	

Table 3. Citations to head and tail articles beyond information behavior.

Periods		Total	Head Articles	Tail Articles	Tail Articles with Impact
Pre E-Research	Core Articles	478	20% (96/478)	80% (382/478)	24.61% (94/382)
	Number of Citations	1,311	88.79% (1,164/1,311)	11.21% (147/1,311)	
	Average Number of Citations	2.74	12.13	0.38	
E-Research Emergent	Core Articles	1,086	20% (217/1,086)	80% (869/1,086)	40.97% (356/869)
	Number of Citations	6,944	86.02% (5,973/6,944)	13.98% (971/6,944)	
	Average Number of Citations	6.39	27.53	1.12	
E-Research Prevalent	Core Articles	4,051	20% (810/4,051)	80% (3,241/4,051)	59.64% (1,933/3,241)
	Number of Citations	42,445	80.03% (33,970/42,445)	19.97% (8,475/42,445)	
	Average Number of Citations	10.48	41.94	2.61	

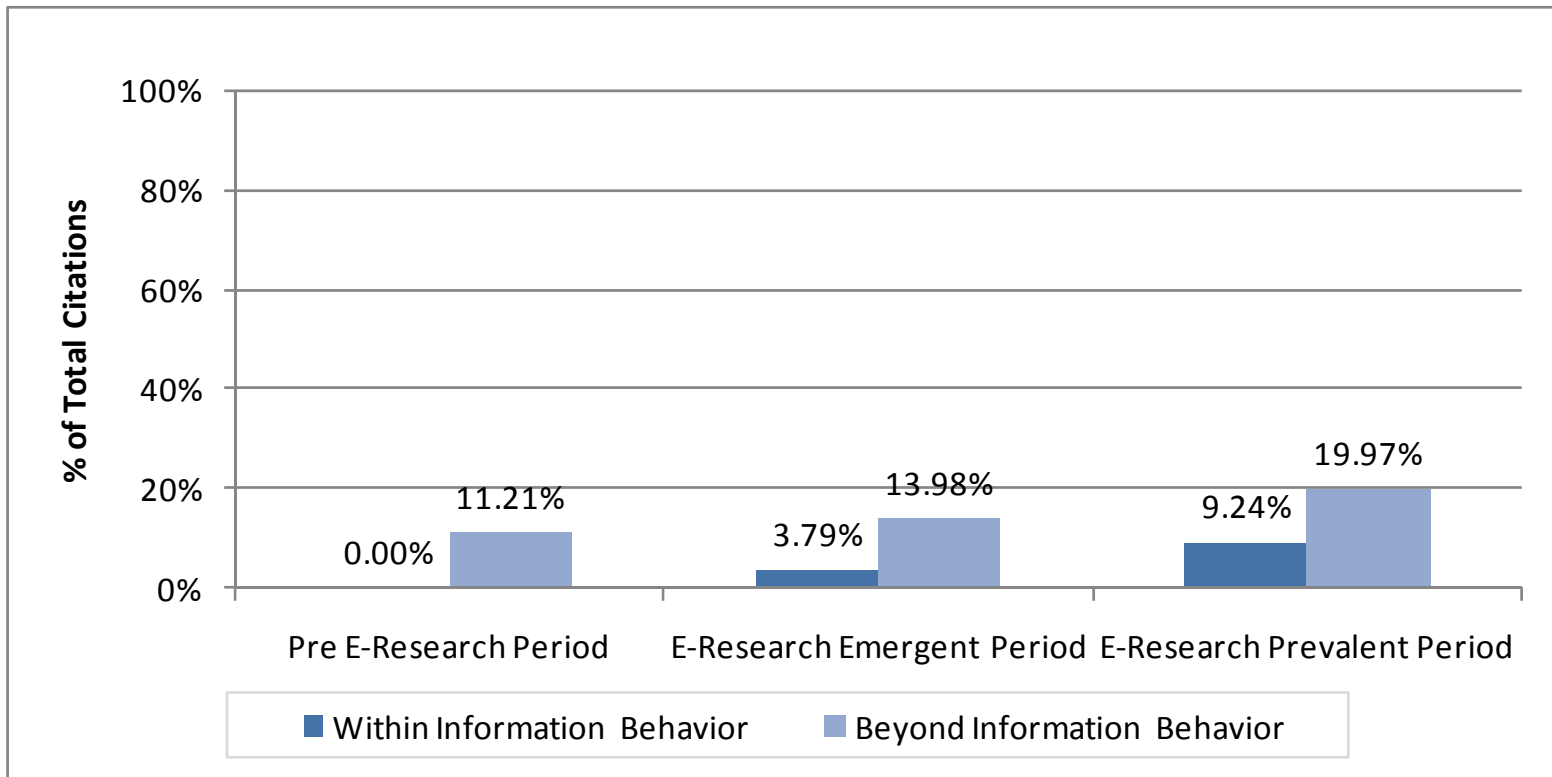


Figure 3. Comparison of growths in proportions of citations to tail articles within and beyond information behavior.

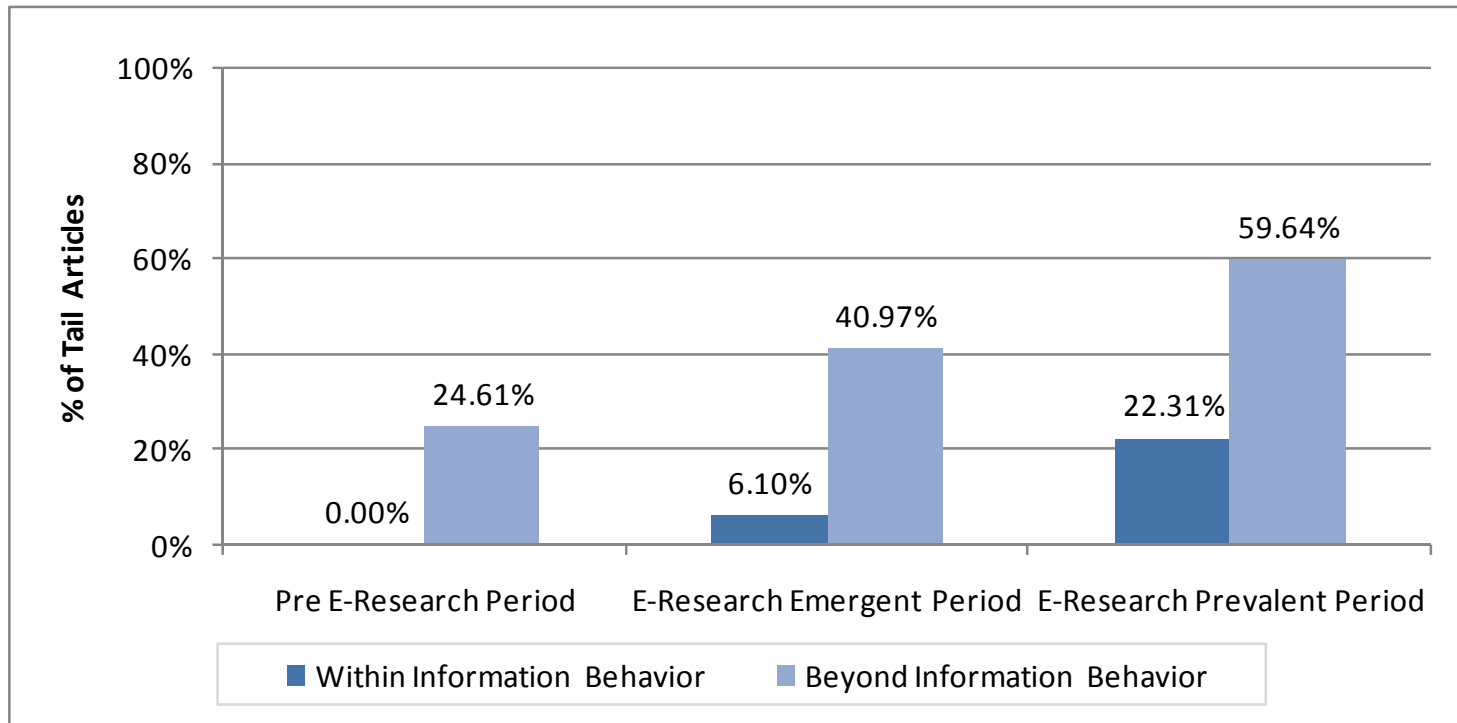


Figure 4. Comparison of growths in proportions of tail articles with impact within and beyond information behavior.

Conclusions and implications

- **The proportion of tail articles that receive citations increase with availability of online articles.**
- **The data of this research supports the Long-tail theory, instead of the Winner-take-all theory.**
- **Relevancy, besides frequency of citation, should also be considered in displaying search results of research articles.**

Future research

- **Comprehensive data set of natural sciences, social sciences and humanities should be analyzed.**
- **Differences between different fields?**

**Thank you for listening.
Questions?**