N-gram and similarity – based discovery of research methodologies, focuses and trends in supply chain management

Shesen Guo[1, a], Ganzhou Zhang[1, b] , Yufei Guo[2, c]

[1] Economics Research, School of Foreign Languages, Qianjiang College, Hangzhou Normal University, Hangzhou, 310018, China.

[2] School of Culture Industries & Management, Xianda College of Economics & Humanities, Shanghai International Studies University, Shanghai 200083, China

[a]email: guoshesen@126.com, [b]email:findsolute@163.com, [c]email: shh_yufeiguo@163.com

Keywords: N-gram; Similarity; Research methodologies; Trends; Supply chain management

Abstract. This work presents an analysis of research methodologies, focuses and trends in supply chain management on the basis of large amounts of data from the Web of Knowledge. Using n-gram and similarity measures, we perform large scale of computing the paper abstracts and author keywords in supply chain management. This work proposes that the two most frequently used research strategies in the research supply chain management are modelling and case study, and this work attempts to provide an overview of changes in trendy ideas or focuses.

Introduction

Quantitative methodologies have been widely used in a variety of areas. Such techniques that are used in economics and management include, for example, analyses of economics journals [1], financial crisis [2], business ethics [3], accounting [4], and corporate governance [5]. In the research of supply chain management (SCM), Giunipero, Hooker, Joseph-Matthews, Yoon, and Brudvig proposed that the quantitative techniques accounted for the largest percentage of research methodologies [6], but their conclusion was based on only 405 articles from 9 journals.

McKinnon stated that in SCM, very limited research that stressed quantitative models could be published in top journals [7]. And he did not provide clear and strong evidence to justify his assertion. To date, no research explores the objective statistics of methodologies used in SCM from the angle of n-gram and similarity techniques based on large samples.

By using n-gram and similarity models, in this work, we attempt to accurately answer the following questions: 1) What are the percentages of quantitative methodologies used in SCM? 2) What are focuses and trends of research expressed in the keywords or terms by the authors? 3) How are research diversity and richness expressed in keywords and terms in SCM research?

Data and methods

The data used for the measurement was retrieved from the Web of Knowledge (Wok) [8]. The databases were Social Sciences Citation Index (SSCI), Conference Proceedings Citation Index-Social Science & Humanities (CPCI-SSH), and Book Citation Index– Social Sciences & Humanities (BKCI-SSH). The time span was from 2005 to May, 2013. The search was done in October 2013.

Terms used for search were based on all those key, practical, plain, straightforward, and laconic ones from the front page of the famous SCM website http://www.logisticsworld.com/.

We searched the keywords in the search field “Topic” in Wok. “Topic” consists of Title, Abstract, Author Keywords, and Keywords Plus. The search syntax we used included wildcard characters to retrieve the largest number of keywords. The search syntax was Topic=(logistics) OR Topic=(freight) OR Topic=("supply chain") OR Topic=(transport*) OR Topic=("physical distribution") OR Topic=(warehous*) OR Topic=(manufactur*) OR Topic=(truck*) OR Topic=(airline*) OR Topic=(maintenance) OR Topic=(rail*) OR Topic=(ship*) OR

Topic=(container*).

The above databases we used under this syntax returned a total of 45,556 records. To process the retrieved data, we built several tools for parsing the records, analyzing Keyword in Context (KWIC) and building n-gram statistical models. In n-gram computations, we used these widely used stop words [9] for unigram analysis. For bigram, trigram and 4-gram analyses, we included the stop words because they were concerned with some meaningful phrases (e.g. “and” in “small and medium enterprises”; “for” in “demand for green products”). Specifically, for meaningful observation of bigrams, we excluded those that were composed of two function words with unidentifiable meanings (e.g. “of the”) and only listed those that contained two content words.

Results and discussion

Of the total 45,556 records, 42,951 records were with paper abstracts, amounting to 94.282%. The analysis of top 20 unigrams and bigrams in 2005-2013 are listed in Table 1.

Table 1: Top 20 unigrams and bigrams in 2005-2013 paper abstracts

Rank	Unigram	Freq	% of unigrams	Bigram	Freq	% of bigrams
1	model	25860	0.365%	supply chain	17604	0.249%
2	supply	23639	0.334%	case study	2881	0.041%
3	chain	19724	0.279%	chain management	2776	0.039%
4	management	17393	0.246%	decision making	2412	0.034%
5	system	15964	0.225%	supply chains	2222	0.031%
6	data	15631	0.221%	design methodology	1883	0.027%
7	manufacturing	15290	0.216%	methodology approach	1880	0.027%
8	performance	15098	0.213%	originality value	1878	0.027%
9	analysis	14285	0.202%	paper presents	1713	0.024%
10	time	13768	0.194%	long term	1635	0.023%
11	development	13088	0.185%	land use	1427	0.020%
12	cost	12572	0.178%	public transport	1427	0.020%
13	transport	12166	0.172%	logistic regression	1402	0.020%
14	industry	11338	0.160%	manufacturing firms	1401	0.020%
15	process	11136	0.157%	practical implications	1389	0.020%
16	service	10624	0.150%	results indicate	1333	0.019%
17	market	10484	0.148%	research limitations	1124	0.016%
18	logistics	10466	0.148%	manufacturing industry	1118	0.016%
19	product	9980	0.141%	limitations implications	1103	0.016%
20	demand	9592	0.135%	paper examines	1075	0.015%

Freq = Frequency

The word *model* is the most frequently used content word in all 2005-2013 abstracts, exceeding considerably supply, chain and management. This indicated that a substantial proportion of publications were employing simplified systems or quantitative methodologies in research. The token data also suggests that researchers were using quantitative methods in their studies. The bigram *case study* is ranked the second with 2,881 times. If the phrase *case study* appeared once in one abstract, 6.708% of the 42,951 records contained the information about such methodology in research. Table 1 clearly shows that modeling and case study are the two most important and frequently used research methodologies in SCM research.

In the 42,951 records, the total occurrences of the 3-word types that contained the word *model* were 30,573, indicating a variety of models used in research. Table 2 lists the top 20 3-word types that contained suffix *model*. Obviously, *Structural equation model* significantly exceeds other statistical models in number.

Table 2: Top 20 3-word types with suffix *model*

Rank	Type	Frequency	% of 30,573
1	structural equation model	538	1.760%
2	logistic regression model	389	1.272%
3	a conceptual model	222	0.726%
4	supply chain model	169	0.553%
5	a mathematical model	148	0.484%
6	integer programming model	138	0.451%
7	discrete choice model	127	0.415%
8	linear programming model	120	0.393%
9	a simulation model	110	0.360%
10	a theoretical model	99	0.324%
11	general equilibrium model	98	0.321%
12	multinomial logit model	86	0.281%
13	the mathematical model	85	0.278%
14	travel demand model	79	0.258%
15	an optimization model	73	0.239%
16	nested logit model	69	0.226%
17	an integrated model	68	0.222%
18	system dynamics model	64	0.209%
19	the theoretical model	62	0.203%
20	mathematical programming model	57	0.186%

In the abstracts，we found that a sentence starting with “ Supply chain management has” or “Supply chain management is” usually represents key concepts, perception, or summary proposed by the authors. For example, “Supply chain management has become a potentially valuable way of securing competitive advantage and improving organization performance since competition is no longer between organizations, but among supply chains.” We extracted all sentences with such prefixes to observe focus changes and trend in SCM research, we divided the 2005-2013 data into three 3-year-period datasets (2005-2007 document, 2008-2010 document and 2011-2013 document) and made comparative similarity analyses based on vector space model (VSM) below [10]:

$$similarity(d_i, d_j) = \frac{\vec{V}(d_i) \bullet \vec{V}(d_j)}{\|d_i\| \|d_j\|}$$

Where similarity (d_i , d_j) is similarity measure of document i and document j. $\vec{V}(d_i) \bullet \vec{V}(d_j)$ stands for the inner product of document *i* and document *j* vectors. $\| d_i \|$ is the norm of document *i* vector, and $\| \mathrm{d}_\mathrm{j} \|$ denotes the norm of document *j* vector. By computing term frequency (tf) of each document and inverse document frequency (idf), we had tf-idf values for the above formula. Table 3 shows the result of similarity comparisons.

Table 3: Similarity measures of documents of 3 periods

Period1	Period2	Similarity (period1, period2)	Year difference between period1 and period2
2005-2007 (65)*	2011-2013 (40)	0.03225	4
2008-2010 (91)	2011-1013 (40)	0.06777	1
2005-2007 (65)	2008-2010 (91)	0.10955	1

* values in parenthesis denote frequencies of *Supply chain management is (has)*

Table 3 shows that there was a larger difference between 2005-2007 and 2011-2013 than that between 2008-2010 and 2011-2013 and there was a much larger difference between 2005-2007 and 2011-2013 than that between 2005-2007 and 2008-2010. Low similarity indicates comparatively dramatic changes in terms used after Supply chain management is (has), implying greater and newer innovations and research interest are introduced. Though there was one year difference both between 2008-2010 and 2011-2013 and there was one year difference between 2005-2007 and 2008-2010, the changes in the contexts after *Supply chain management is (has)* in 2011-2013 were comparatively and immediately more noticeable.

We computed all available author keywords of the papers to track the research focuses in SCM. Table 4 lists the top 30 most frequently used author keywords from 2005 to 2013. *China*, as an emerging and developing economy, is one of the hot topics in the whole investigative period. *Innovation* is a constant and central theme in SCM. *Simulation* is closely related to modeling, indicating researchers are building or using models of real systems for experiments or solutions. *Sustainability* is capturing increasing attention in the research, relating to considerations of environmental, social, and economic factors in order to preserve continued viability. *Optimization* is a popular theme throughout the whole period, concerned with reducing costs and improving performance. *Reverse logistics* declines in rank, from 9th position in 2005-2007 to 17th position in 2008-2010 and 21st in 2011-2013. *Scheduling* is moving up its positions to reach 13th in 2011-2013. *Logistic regression* is drawing increasing attention as a statistical analysis. It is especially striking that *climate change* rose to prominence from 112th place in 2005-2007 to 39th in 2008-2010 and 12th in 2011-2013. There were 24 documents in 2005-2007 that contained *climate change* in author keywords, 66 in 2008-2010, and 103 in 2011-2013. It is likely that *climate change* may continue to rise or maintain stable in author keywords rank in future SCM research.

Table 4: Top 30 most frequently used author keywords from 2005 to 2013 (ascending in 2005-2013)

Author	2005-2007		2008-2010		2011-2013		2005-2013	
keywords	Rank	PAK	Rank	PAK	Rank	PAK	Rank	PAK
supply chain management	1	1.319%	1	1.276%	1	0.945%	**1**	1.175%
supply chain	2	1.064%	2	0.886%	2	0.475%	**2**	0.791%
China	5	0.282%	4	0.383%	3	0.413%	**3**	0.368%
logistics	3	0.438%	3	0.396%	7	0.249%	**4**	0.357%
innovation	4	0.347%	5	0.345%	4	0.338%	**5**	0.343%
simulation	7	0.274%	6	0.288%	8	0.218%	**6**	0.261%
transportation	8	0.268%	7	0.236%	6	0.261%	**7**	0.252%
manufacturing	6	0.276%	8	0.233%	9	0.214%	**8**	0.237%
inventory	10	0.235%	10	0.226%	11	0.201%	**9**	0.219%
sustainability	32	0.125%	12	0.176%	5	0.263%	**10**	0.193%
performance	16	0.159%	9	0.233%	14	0.159%	**11**	0.190%
game theory	20	0.146%	11	0.176%	10	0.212%	**12**	0.181%
optimization	11	0.227%	13	0.175%	17	0.150%	**13**	0.179%
reverse logistics	9	0.261%	17	0.169%	21	0.133%	**14**	0.179%
pricing	12	0.188%	14	0.172%	15	0.159%	**15**	0.172%
productivity	13	0.175%	15	0.170%	16	0.156%	**16**	0.166%
outsourcing	15	0.159%	16	0.169%	22	0.128%	**17**	0.153%
scheduling	28	0.130%	19	0.145%	13	0.161%	**18**	0.147%
transport	17	0.159%	24	0.126%	19	0.150%	**19**	0.142%
case study	18	0.156%	20	0.142%	26	0.122%	**20**	0.139%
sustainable development	21	0.141%	18	0.146%	30	0.113%	**21**	0.134%
knowledge management	14	0.164%	29	0.114%	28	0.118%	**22**	0.127%
trust	33	0.123%	22	0.130%	25	0.124%	**23**	0.126%
climate change	112	0.063%	39	0.099%	12	0.193%	**24**	0.122%
globalization	31	0.125%	23	0.127%	36	0.098%	**25**	0.117%
logistic regression	37	0.112%	37	0.100%	20	0.139%	**26**	0.116%
competition	29	0.128%	26	0.118%	32	0.103%	**27**	0.115%
heuristics	26	0.136%	34	0.108%	31	0.109%	**28**	0.115%
uncertainty	34	0.123%	38	0.100%	23	0.126%	**29**	0.114%
public transport	46	0.099%	46	0.093%	18	0.150%	**30**	0.114%

PAK: Percentage of total number of author keywords

The similarity measures of author keywords are shown in Table 5. They indicate that trendy ideas or concepts underwent gradual and continued change over time. One or two year difference between 2010-2012, 2011-2012, 2009-2010, 2010-2011, and 2005-2006 matched with comparatively higher degree of similarity (first 5 rows). Additionally, longer span of time (2006-2013, 2005-2013, 2007-2010, 2007-2012, 2007-2011, and 2007-2013) displayed lower degree of similarity, which may suggest comparatively larger changes in research interest or focus. 2007 was a special year. Author keywords in 2007 were comparatively more different from those before and after 2007 (last 5 rows), though 2006, 2007 and 2008 were three consecutive observation years. This possibly suggests that comparatively lower proportion of new ideas, concepts, theories or focus of research introduced in 2007 were extending into the subsequent years.

Table 5: Author keywords similarity comparison from 2005 to 2013

Year1	Year2	Similarity (Year1,Year2)	Year difference between Year1 and Year2
2010	2012	0.04	2
2011	2012	0.03179	1
2009	2010	0.02984	1
2010	2011	0.02822	1
2005	2006	0.02724	1
2008	2012	0.02168	4
2009	2011	0.02117	2
2009	2012	0.018	3
2008	2009	0.01777	1
2008	2011	0.01569	3
2005	2008	0.01489	3
2012	2013	0.01434	1
2008	2010	0.01382	2
2006	2008	0.01246	2
2005	2012	0.01207	7
2005	2009	0.01188	4
2005	2011	0.01173	6
2006	2012	0.01161	6
2010	2013	0.01132	3
2006	2010	0.01076	4
2006	2011	0.01009	5
2009	2013	0.01008	4
2005	2010	0.00994	5
2011	2013	0.00937	2
2008	2013	0.00688	5
2005	2007	0.00687	2
2006	2013	0.0067	7
2007	2008	0.00606	1
2006	2009	0.00605	3
2005	2013	0.00526	8
2007	2010	0.00523	3
2006	2007	0.00514	1
2007	2012	0.00442	5
2007	2011	0.00404	4
2007	2009	0.00392	2
2007	2013	0.00304	6

Conclusion

We have presented an analysis of research focuses and trends in SCM on the basis of n-gram and similarity techniques. It has been found that modeling, empirical or case studies are the most important research methodologies in SCM. The structural equation model is the most frequently used statistical model. The period from 2005 to 2010 experienced a gradual introduction of new concepts. The 2010-2013 period witnessed dramatic and more introductions of new ideas in comparison with previous periods. The keywords used for the research in this period were largely hapax legomena, demonstrating immense diversity. *Innovation, simulation (modeling), sustainability, optimization, performance, system* were some central and recurring themes. *China* and the *USA* were the two countries that were capturing attention of researchers. *Climate change* was the term that rose dramatically in keywords rankings. In the future, the position of this term may continue to rise in SCM research.

References

[1] T. F. Frandsen. Journal interaction: A bibliometric analysis of economics journals [J]. Journal of Documentation, 2005 61(3) 385 - 401.

[2] Chong-Chuo Chang, Yuh-Shan Ho. Bibliometric analysis of financial crisis research. African Journal of Business Management [J], 2010 4(18) 3898-3910.

[3] Ö. Ö. Uysal. 2010. Business ethics research with an accounting focus: A bibliometric analysis from 1988 to 2007 [J]. Journal of Business Ethics, 2010 93(1) 137-160.

[4] A. Just, M. Meyer, E. Perrey, U. Schäffer. Accounting as a normal science? An empirical investigation of the intellectual structure of accounting research [C]. Annual Meeting of the American Accounting Association. July 31- August 4, 2010, San Francisco, CA.

[5] Chiung-Yao Huang, Yuh-Shan Ho. Historical research on corporate governance: A bibliometric analysis [J] African Journal of Business Management, 2011 5(2) 276-284.

[6] L.C. Giunipero, R. E. Hooker, S. Joseph-Matthews, T. E.Yoon, S. Brudvig. A decade of SCM literature: past, present and future implications [J]. Journal of Supply Chain Management, 2008 44(4) 66-86.

[7] A.C. McKinnon. Starry-eyed: journal rankings and the future of logistics research [J]. International Journal of Physical Distribution and Logistics Management, 2013 43(1) 6-17.

[8] Web of Knowledge (Wok). Web of Knowledge – The facts. Retrieved from http://webofknowledge.com/

[9] Ranks.nl Stopword list. 2013. Ranks – Resources- Stopword list. Retrieved from http://www.ranks.nl/resources/stopwords.html

[10] G. Salton. Automatic text processing: The transformation, analysis and retrieval of information by computer [M]. Reading, MA: Addison-Wesley, 1989.