Text Mining the CMG Archives

Richard Gimarc
rgimarc@featherfall.com

April 17, 2019
Southwest CMG

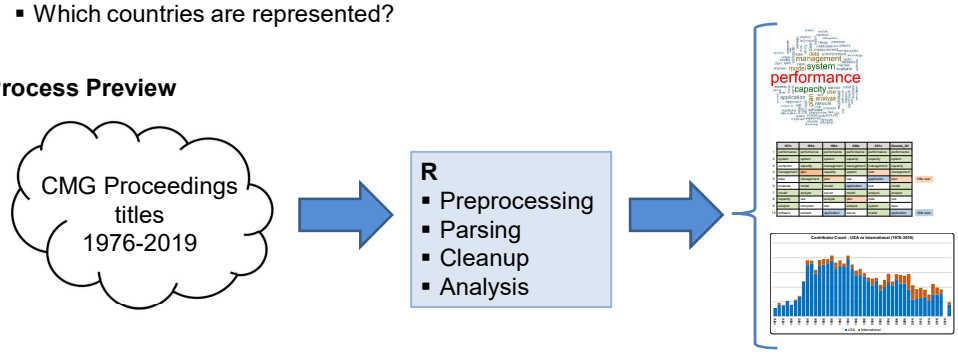
Text Mining The CMG Archives

Given: Titles & authors from CMG conferences 1976-2019

What can we learn about CMG?

- Do the words used in paper/presentation titles tell us anything about CMG?
- Has our choice of words changed over the years?
- Who is contributing content to CMG?
- Which countries are represented?

Process Preview

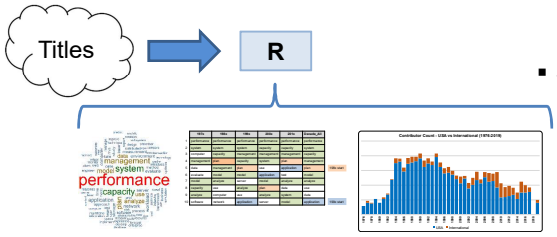


1

Text Mining Process – Titles Processing & Cleanup

What do we include?

- Include papers & presentations from annual CMG conference (1976-2019)
- Only include downloadable content (e.g., PDF) from conference proceedings
- Exclude pre-conference workshops
- Each CMG-T tutorial counted as +1

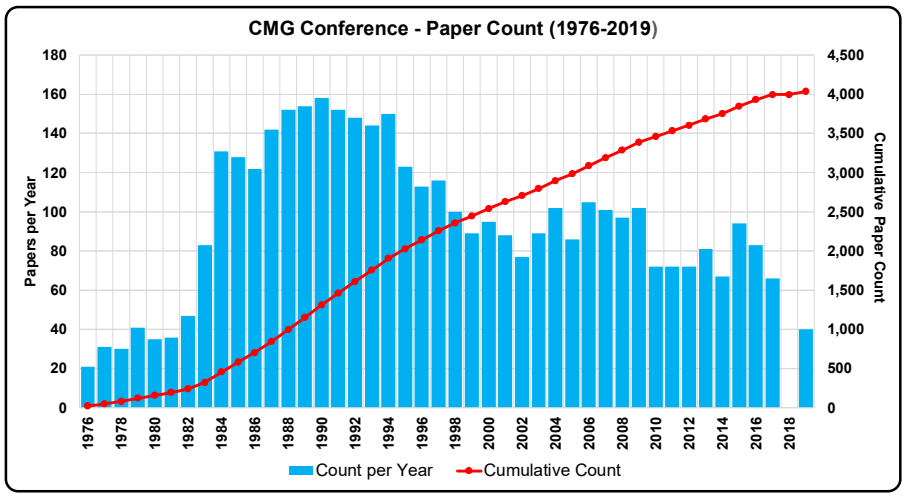


Preprocessing, Parsing & Cleanup

- What do we identify for each paper/presentation?
  - **Title** from PDF rather that what was in the agenda
  - Verify title & complete set of **authors** on each paper/presentation
  - Identify **country** for each author
- Remove stop words (frequent but provide little information)
  - Examples: *the*, *is*, *at*, *which*, and *on*.
- Standardize words
  - "*modelling*" → "*model*" & "*modeling*" → "*model*"
- Standardize names
  - " *Amy C. Spellmann*" → " *Amy Spellmann*"

2

CMG Archive – CMG Conference Proceedings (input)



Total papers/presentations 3,964

3

Word Cloud

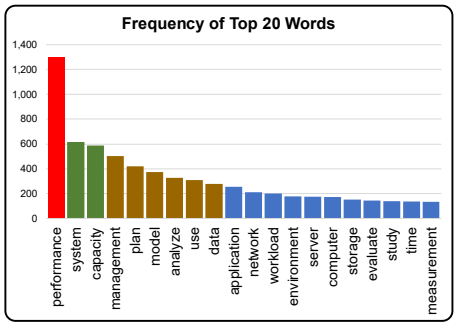
Word cloud created from the titles of papers/presentations 1976-2019

Word Cloud: an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

4

Word Cloud

Verification



5

Has our choice of words changed over the years? (nope)

	197x	198x	199x	200x	201x	Decade_All	
1	performance	performance	performance	performance	performance	performance	Consistency
2	system	system	system	capacity	capacity	system	
3	computer	capacity	management	management	management	capacity	
4	management	plan	capacity	system	plan	management	
5	data	management	plan	use	application	plan	198x start
6	evaluate	model	model	application	test	model	
7	model	analyze	server	model	analyze	analyze	
8	capacity	use	analyze	plan	data	use	
9	analyze	computer	use	analyze	system	data	
10	software	network	application	server	model	application	199x start

6

Who is contributing content to CMG?

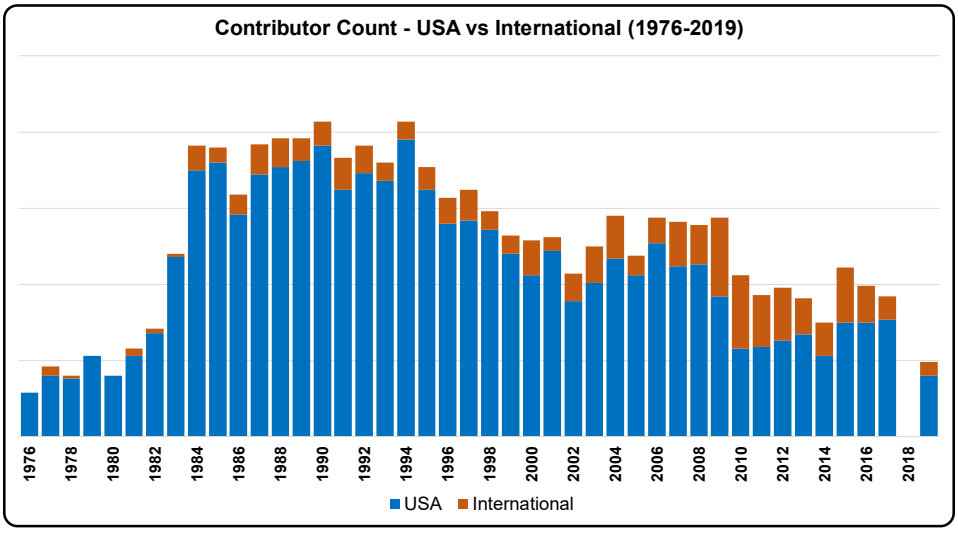
Top 10 Authors	Count
Bernie Domanski	43
Connie U. Smith	43
Jeffrey P. Buzen	38
H. Pat Artis	37
Michael Salsburg	37
Yiping Ding	37
Daniel A. Menascé	35
Bruce McNutt	33
Mark Friedman	32
T. Leo Lo	31

2,548 Individual Contributors (1976-2019)

Today's presenters	Count
Richard Gimarc	29
Amy Spellmann	19
Barry Merrill	8
Ben Davies	4
Brent Phillips	1

7

Is CMG an international organization?



8

Individual contributions from 35 countries

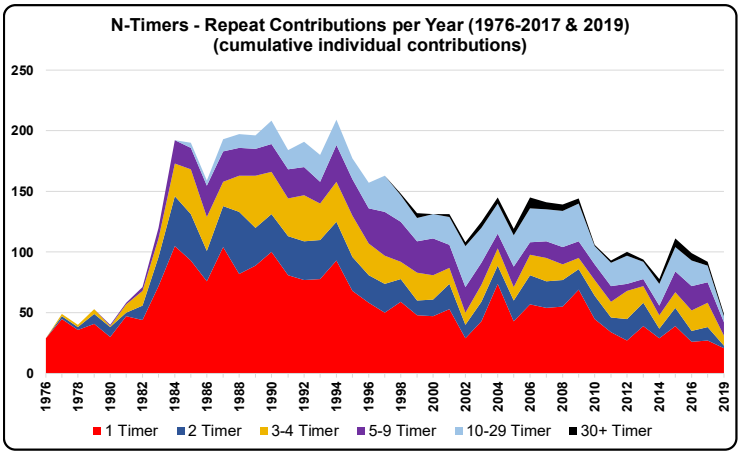
Author	Count
USA	4,711
Canada	133
India	133
England	118
Italy	72
The Netherlands	69
Australia	37
Brazil	25
South Korea	25
Japan	23
Germany	22
China	21

Author	Count
South Africa	8
France	7
Spain	7
Russia	6
Scotland	6
Ireland	5
Belgium	4
Denmark	4
Norway	4
Puerto Rico	4
Austria	3
Switzerland	3

Author	Count
Saudi Arabia	2
Sweden	2
Argentina	1
Finland	1
Iceland	1
Israel	1
Pakistan	1
Portugal	1
Singapore	1
Uruguay	1
Venezuela	1

9

Do authors return to make additional contributions?



Definition

- An "***N-Timer***" is someone who has made N contributions to the annual conference
- The set of "***N-Timers***" changes for each succeeding year of the conference

Observations:

- 1-Timers are the largest single class
- Most recent conference has a noticeably smaller number of repeat contributors

- Everyone starts as a 1-Timer
- When you present your 2nd paper, you are promoted to a 2-Timer.
- Rinse & repeat

10

Interesting Observations

Shortest title	"***8½***" - 1993 paper by James McGilliard at FEDSIM. Title is a reference to a 1963 movie by Federico Fellini.
Largest title *(magnitude)*	"***40,153,273,652***" - 1998 paper by Chuck Hopf and Barry Merrill at Merrill Consultants. This is the number of bytes of SMF data produced by a single complex in a single day.
One word title	"***MIPS***" - 2000 paper by Tim Follen at Medical Mutual of Ohio "***blockchain***" - 2019 presentation by John deVadoss at NEO.org
First use of an email address	**raghu%deceat.dec.com** and **biswas%xanadu.dec.com** - 1990 paper by Melur K. Raghuraman & Prabuddha Biswas from DEC

11

Text Mining The CMG Archives

What did we learn about CMG? (answering the questions)

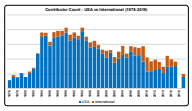
1) **Do the words used in paper/presentation titles tell us anything about CMG?**
   - Developed word cloud from conference paper/presentation titles (3,964)
   - High frequency words support CMG Purpose (Bylaws, Article 2):

     Foster research and development, and the exchange and public dissemination of data pertaining to computer measurement, computer management, and computer performance evaluation, and underlying computer science.

2) **Has our choice of words changed over the years?**
   - Not much; 6 words in top 10 list for all 5 decades & overall
   - #1 word is "*performance*"

3) **Who is contributing content to CMG?**
   - 2,548 contributors
   - Examined shape of N-Timer participation

Top 10 Authors	Count
Bernie Domanski	43
Connie U. Smith	43
Jeffrey P. Buzen	38
H. Pat Artis	37
Michael Salsburg	37
Yiping Ding	37
Daniel A. Menascé	35
Bruce McNutt	33
Mark Friedman	32
T. Leo Lo	31



4) **Which countries are represented?**
   - 35 different countries

12

Text Mining the CMG Archives

Richard Gimarc
rgimarc@featherfall.com

April 17, 2019
Southwest CMG