A Bibliographic Analysis of Literature on Consumer Health Information and Libraries

Pamela L. Kemp

May 2, 2020

Topic Description

The National Institute of Medicine defines "Consumer Health Informatics" as a field in information science concerned with the analysis and dissemination through the application of computers of information designed with multiple consumers or individual patients as the intended audience (*Consumer Health Informatics*, 2018). "Consumer Health Information", a by-product of consumer health informatics, is the information intended for potential users of medical and healthcare services, with an emphasis on self-care and preventive approaches as well as information for community-wide dissemination and use (*Consumer Health Information*, 2005) .

As Demeris (Demiris, 2016) states, as early as the 1990s scientific literature in the field of consumer health informatics primarily focused on the quality of information that consumers were able to access from libraries, mass media, and the early versions of web-based health information, which would prepare them for a clinical encounter or help them better understand a disease or treatment plan.

This paper is a bibliographic analysis of a representative sample of the currently available literature on Consumer Health Information (CHI) and Libraries. The paper identifies

the prevalent themes in the study and research of the role of libraries in the development and delivery of Consumer Health Information.

I selected this topic because of my interest in this area of librarianship. The results of this analysis will contribute to the development of a controlled vocabulary of terms which will facilitate future information searches and enable me to better assist others conducting research in this area.

Research Question

The goal of this analysis is to answer the following questions:

1. What are the most prevalent themes in the literature on this topic?
2. What are keywords that can be used for information searches on this topic?
3. Who are the most frequently cited authors on this topic and what are their areas of focus?
4. Are there any emerging trends in the literature on this topic?

Data Description

Data for this paper was collected from two different sources for the purposes of conducting both Topic Modeling as well as Bibliometric Analysis in R and VOSviewer.

Phase I

For the Topic Modeling Analysis, I conducted a search in EBSCOhost Academic Search Complete using the query, SU(Libraries) AND TX("Consumer Health Information") Published Date: 20000101-20201231. This yielded 457 results which I exported to an XML file

I used the R xml2 package to import and read the XML file data and save a subset to a dataframe comprised of 440 objects and two variables. The first variable in the dataframe was an integer serving as a document identifier, the second was the document abstract. These results were written to comma delimited file. I then imported the comma delimited file and used the R SnowballC and tm packages to create a corpus, remove unwanted characters, perform tokenization, stopword elimination, and stemming. Next, I created a document term matrix and frequency table which contained 3,558 unique terms. The document term matrix was used to

conduct topic modeling using a Latent Dirichlet Allocation (LDA) and Gibbs sampling. I will discuss this in the next section under Data Analysis and Results The term table was sorted in order of decreasing frequency and saved to a comma delimited file. The following table lists the top 50 terms in descending order of frequency.

Top 50 Terms by Frequency

	Term	Frequency		Term	Frequency
1	health	1143	26	nation	129
2	librari	978	27	program	128
3	inform	875	28	site	128
4	librarian	354	29	medicin	119
5	consum	294	30	present	118
6	public	284	31	role	118
7	medic	264	32	discuss	113
8	servic	252	33	literaci	111
9	provid	251	34	communiti	110
10	resourc	228	35	associ	108
11	articl	225	36	profession	106
12	develop	165	37	review	106
13	patient	164	38	onlin	99
14	research	160	39	refer	98
15	need	152	40	staff	95
16	web	144	41	evalu	91
17	scienc	140	42	issu	90
18	project	136	43	support	89
19	hospit	136	44	train	89
20	result	135	45	search	87
21	studi	134	46	user	87
22	access	133	47	outreach	86
23	educ	132	48	method	85
24	care	131	49	survey	84
25	includ	130	50	internet	80

Phases II and III

Since the search in EBSCOhost Academic Search Complete resulted in less than a sample size of 500 documents, I conducted a search in SCOPUS using the query:

"Consumer Health Information" AND libraries AND (LIMIT-TO (PUBYEAR , 2010 – 2020)) AND (LIMIT-TO (LANGUAGE , "English")). This yielded 1,588 documents.

I exported the bibliographic data in both bibtext for bibliometric analysis in R, and comma delimited format for analysis in VOSviewer.

Data Analysis and Results

Phase I – LDA Topic Modeling

In this phase, I used the R Topicsmodel package to perform a Latent Dirichlet Analysis on the document abstracts which I collected from EBSCOhost Academic Search Complete. As discussed during the course lecture this is a good method for extracting document topics from a large unstructured collection of documents, discovering patterns between documents, and discovering the probability of the distribution of words within topics (Joo, 2020, p. 28). Topic Modeling has many practical applications for library science including information retrieval, content-based recommendation systems, document summary, social media mining for teaching information literacy, and bibliometrics (Joo, 2020). Latent Dirichlet Analysis with Gibbs Sampling is essentially a statistical method for calculating the probability of the distribution of terms within documents and based on the frequency of those terms, creating clusters which define topics. Doll (Doll, 2019) describes this as reverse engineering a topic. Instead of beginning with a topic and discovering which terms belong to the topic, you discover the terms, the frequency of those terms in similar documents, and based upon frequency and similarity, define topics.

I initially ran a LDA analysis to identify 20 topics and 10 terms within each topic. However, the results appeared to be too fragmented. I then opted to identify 15 topics and the top 12 terms within each topic. The results are displayed in Appendix I. As you can see the R program does not assign names but only numbers to the topic. Therefore, the topics are still left to subjective interpretation. I hypothesized that the topics could be identified as: 1.) CHI resource formats; 2.) the need for access to CHI; 3.) how libraries provide training to their staffs and/or the public; 4.) CHI resources referred by clinics; 5.) online CHI resources; 6.) programs conducted by the National Libraries of Medicine; 7.) the role of medical or academic librarians;

8.) medical education and research; 9.) health information literacy; 10.) study and evaluation methods; 11.) articles published in journals; 12.) the Medical Library Association; 13.) digital library resources; 14.) community interaction; and. 15.) searching biomedical resources. Appendix II displays a sample of the topic frequency per document output. It shows that Topic I has a 2.45% frequency of appearing in document one. The table below is a sample of the document to topic analysis. This shows that the main topic for document 1 is topic 3, which I identified as "how libraries provide training to their staffs and/or the public". Appendix II shows that topic 3 occurs with a 17.6% frequency within document 1.

.

Document to Topic Match

doc_id	Topic
1	3
2	6
3	5
4	9
5	10
6	10
7	3
8	9
9	9
10	3

However, due to the subjective nature of the topic labeling, I wanted to conduct other forms of bibliographic analysis.

Phase II – Bibliometric Analysis in R

As explained in the course lecture, bbibliometrics uses quantitative methods, such as statistics and mathematical analysis to examine documents for the correlation that may be derived or inferred in relation to the production, manipulation, or redistribution of information (Joo, 2020). As van Eck and Waltman (van Eck & Waltman, 2014) explain there are three popular approaches for visualizing bibliometric data: distance-based, graph based, and timeline based.

In this phase I use the R bibliometrix package to analyze the data collected from SCOPUS. This analysis identified that the 1,588 were derived from 675 sources. There are:

4583 authors, 2.89 authors per document, 5575 keywords Plus (ID), 3024 author's keywords (DE), and 12.06 average citations per documents. The full summary of the results can be seen in Appendix III. See the visualizations for this data in the following section on Data Visualizations.

Phase III – Bibliometric Analysis in VOSviewer

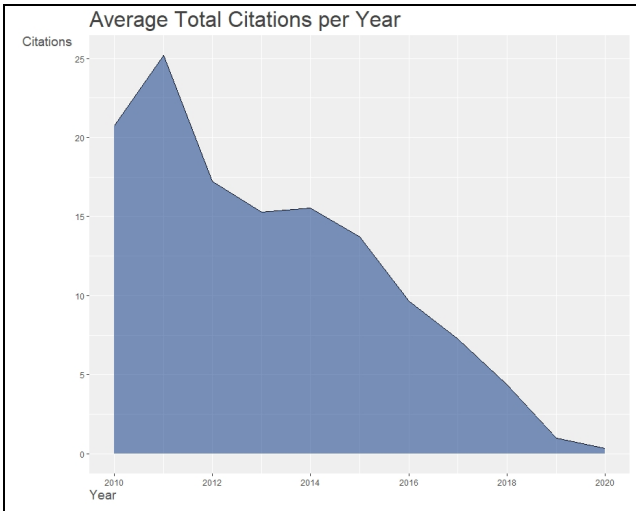
Due to the element of subjectivity in the LDA method of topic identification, I decided to conduct a keyword co-occurrence analysis in VOSviewer. For this analysis I used the same set of SCOPUS data used to conduct the bibliometric analysis in R. VOSviewer identified 7,524 keywords. I limited the visualization to keywords that appeared more than 30 times, 138 terms met that threshold. The resulted analysis identified five clusters and 8030 links, with a total link strength of 84,669.

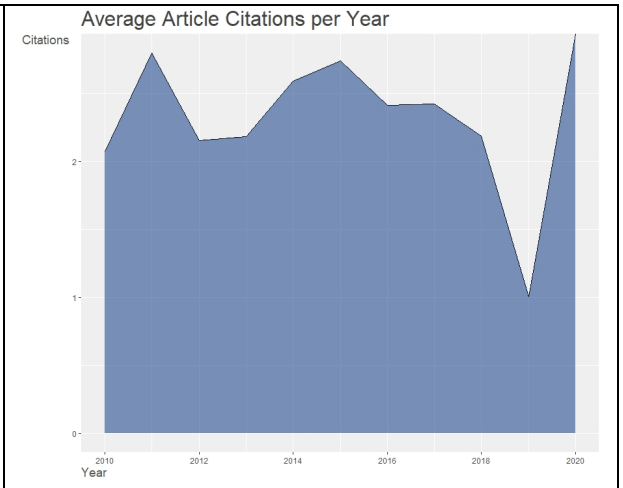
VOSviewer takes a distance-based approach to visualizing bibliometric networks. As van Eck (van Eck & Waltman, 2014) explains a bibliometric network consists of nodes and edges, with the edges indicating the relations between pairs of nodes. Since bibliometric networks are usually weighted networks, the visualizations in the next section show edges indicating not only a relation between two nodes but also the strength of the relation.

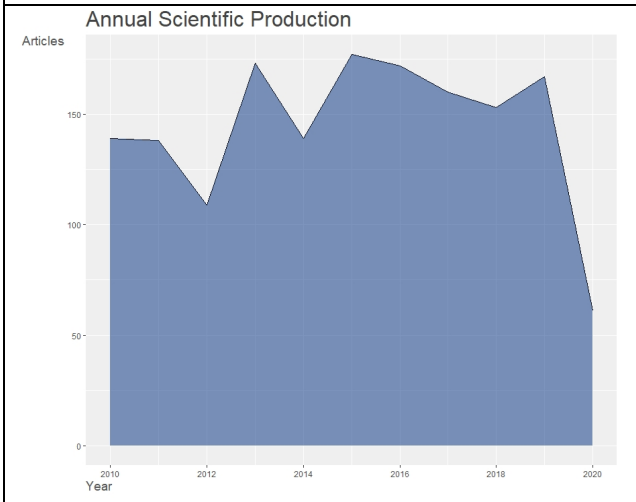
Data Visualization

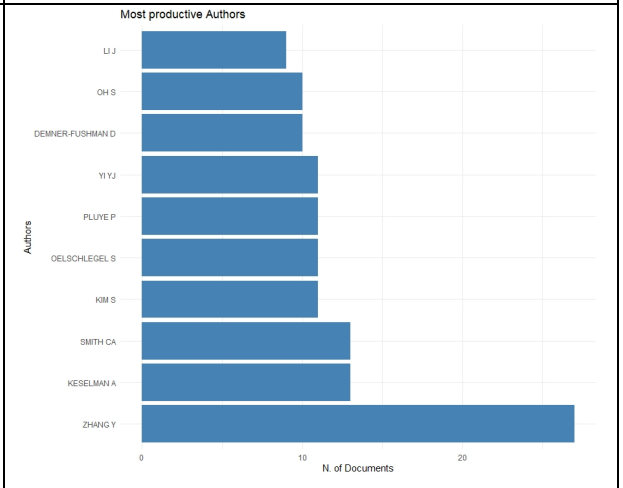
Bibliometric Network Plots in R

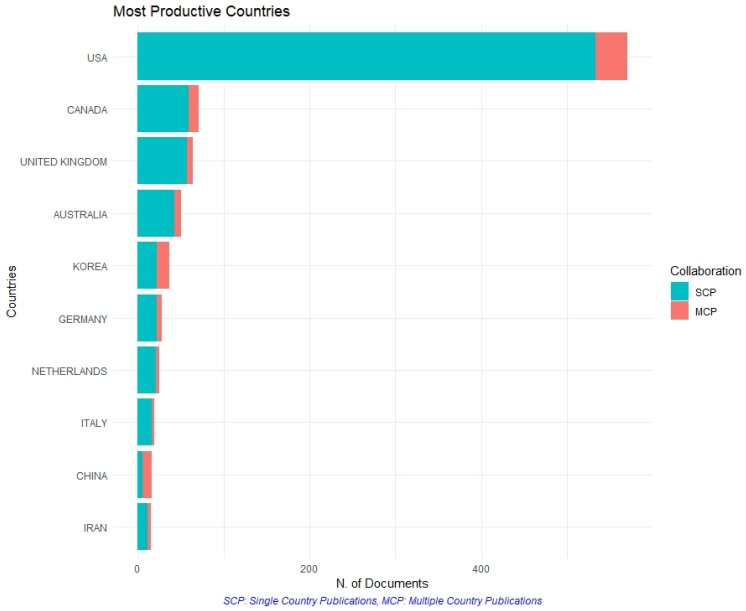
The following visualizations reflect the bibliometric analysis summary in Appendix III


Average Total Citations per Year
Citations
25
20
15
10
5
0
2010
2012
2014
2016
2018
2020
Year


Average Article Citations per Year
Citations
2
1
0
2010
2012
2014
2016
2018
2020
Year

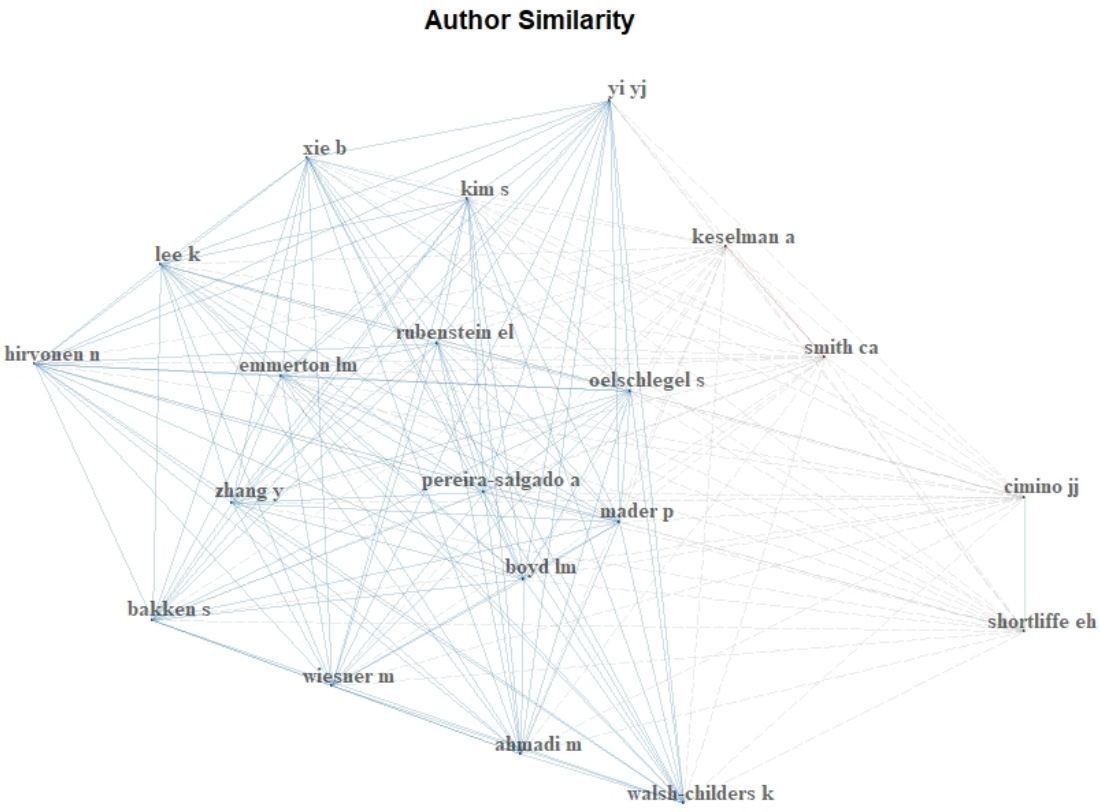

Annual Scientific Production
Articles
150
100
50
0
2010
2012
2014
2016
2018
2020
Year


Most productive Authors
LI J
OH S
DEMNER-FUSHMAN D
YI YJ
PLUYE P
OELSCHLEGEL S
KIM S
SMITH CA
KESELMAN A
ZHANG Y
Authors
0
10
20
N. of Documents

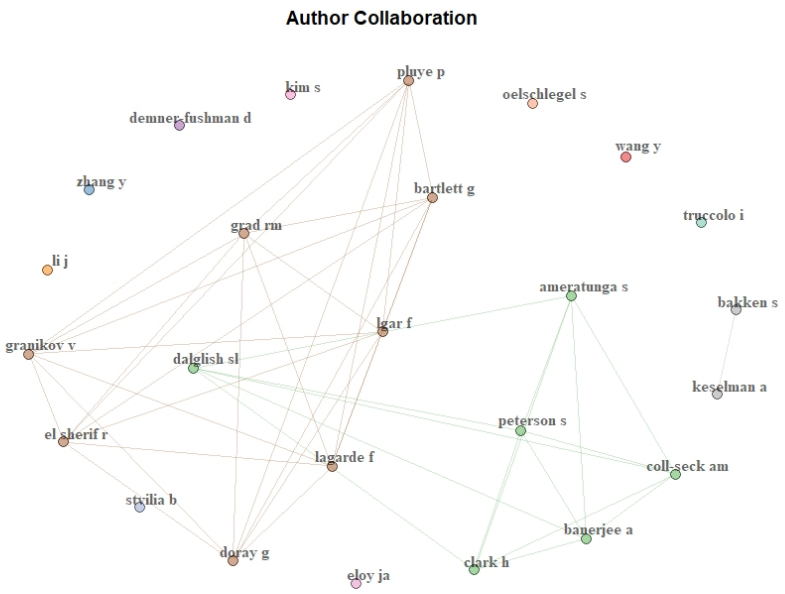

Most Productive Countries
USA
CANADA
UNITED KINGDOM
AUSTRALIA
KOREA
GERMANY
NETHERLANDS
ITALY
CHINA
IRAN
Countries
0
200
400
N. of Documents
Collaboration
SCP
MCP
SCP: Single Country Publications, MCP: Multiple Country Publications

As cited by van Eck and Waltman, the most used techniques for creating graph-based visualizations of bibliometric networks is the graph drawing algorithm of Kamada and Kawai (1989), the algorithm of Fruchterman and Reingold (1991), and the standard circle type.

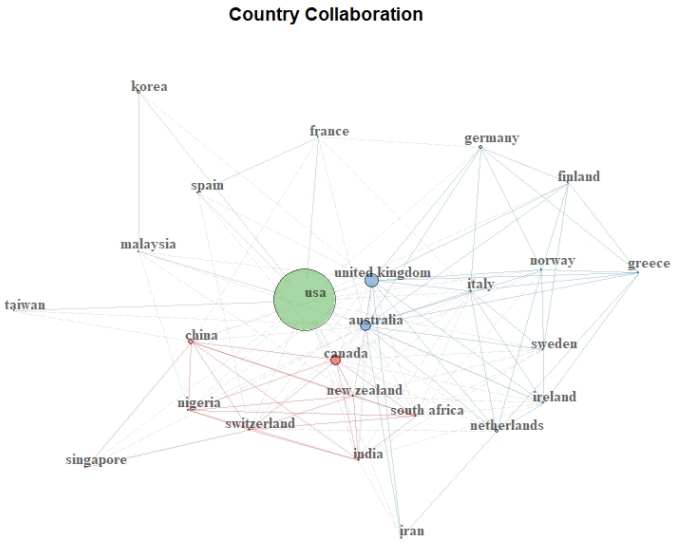
A. This visualization is an author similarity plot using Salton's similarity index and the Fruchterman algorithm



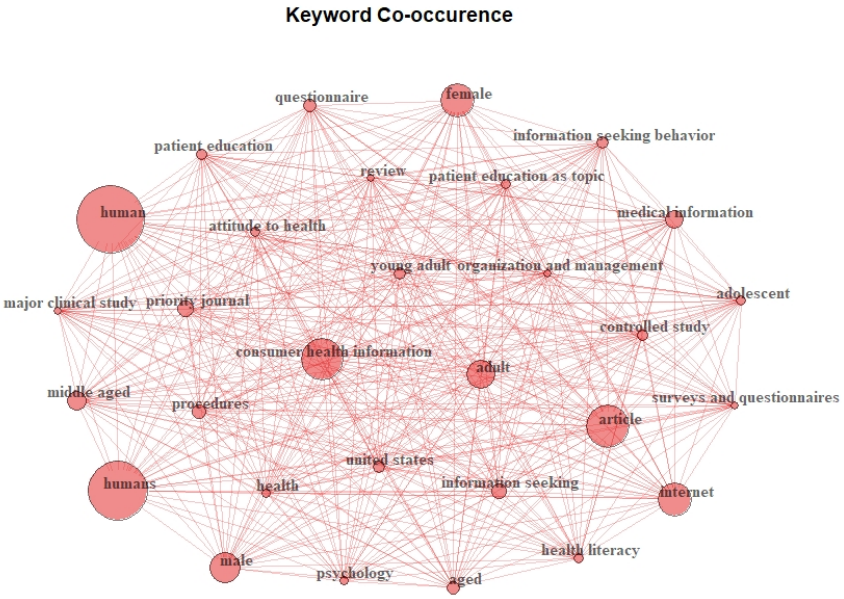
B. The following visualization reflects the author collaboration using the Kamada algorithm



C. The following visualization reflects collaborations by country

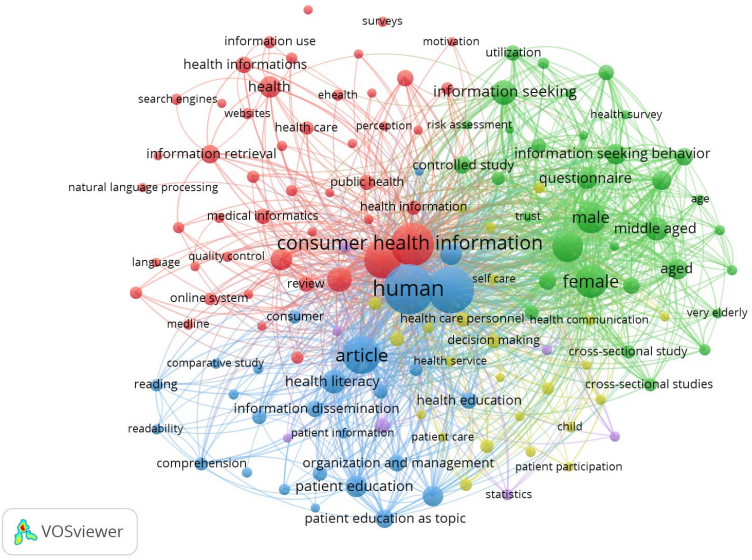


D. This visualization reflects a keyword occurrence

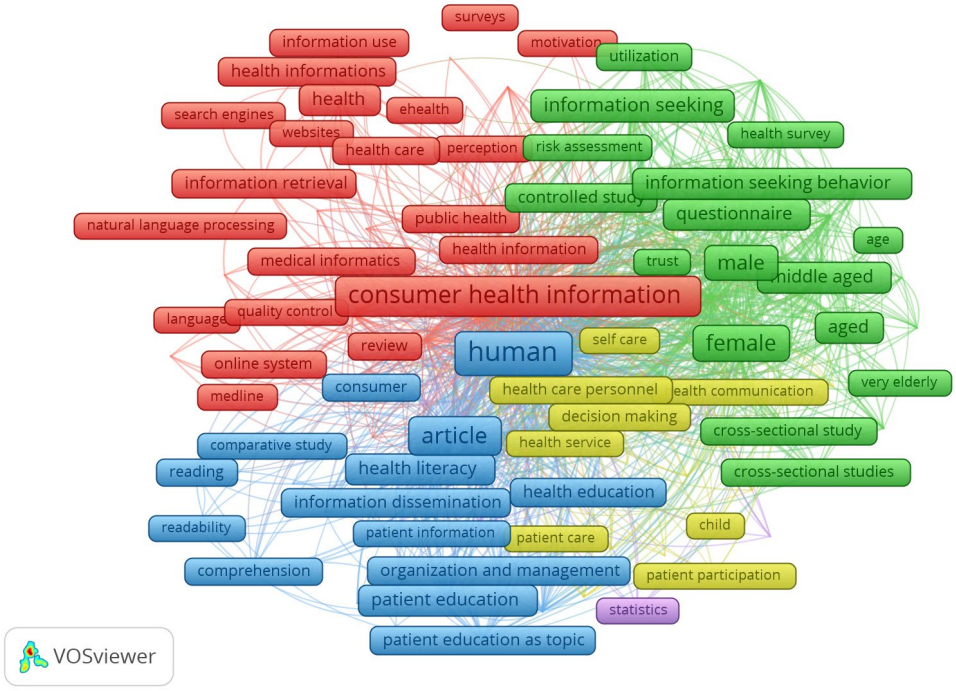


Bibliometric Network Plots in VOSviewer

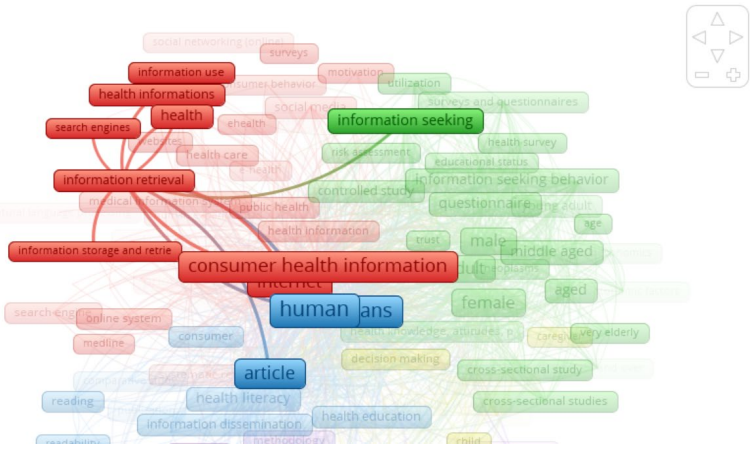
This is an overall visualization of keyword co-occurrence using circle nodes.



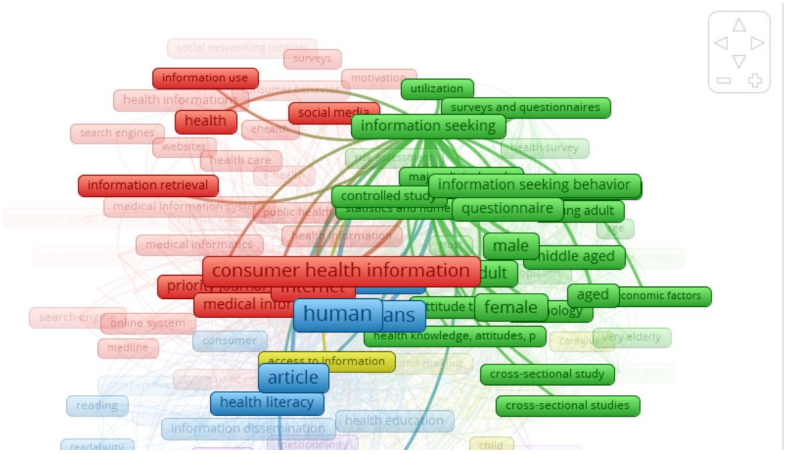
This is a visualization of keyword co-occurrence using frames



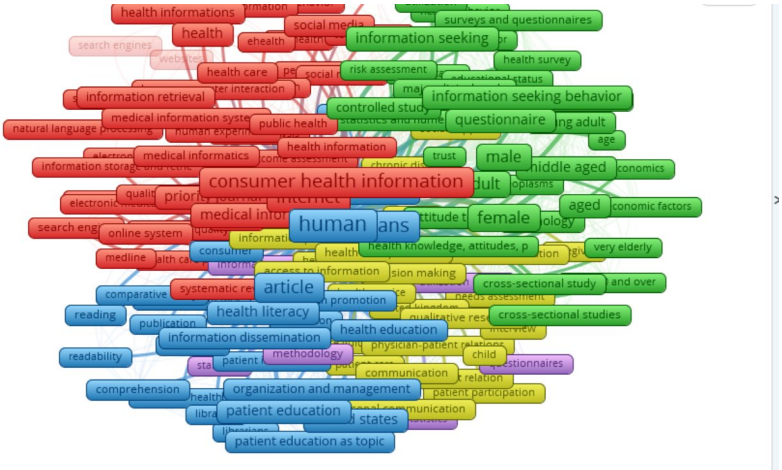
Cluster 1 is represented in red and the main keyword is “customer health information” which 137 links, 7174 total strength and 659 occurrences. Other keywords in this cluster include terms like health information, medical informatics, information retrieval, search engines, websites, online system, Medline and public health.



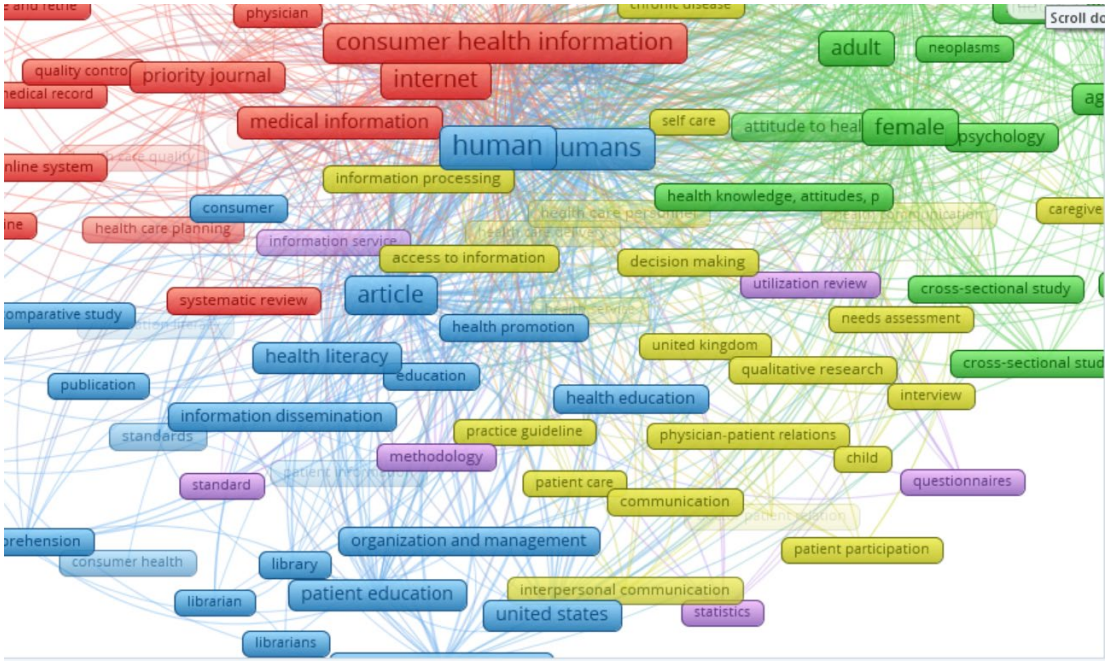
Cluster 2 is represented in green. The main theme of this cluster is patient demographics, behaviors and evaluation methods. The most prevalent keywords are demographic identifiers like adult with 338 occurrences, male with 340 occurrences, and female with 381 occurrences, all with 137 total links.



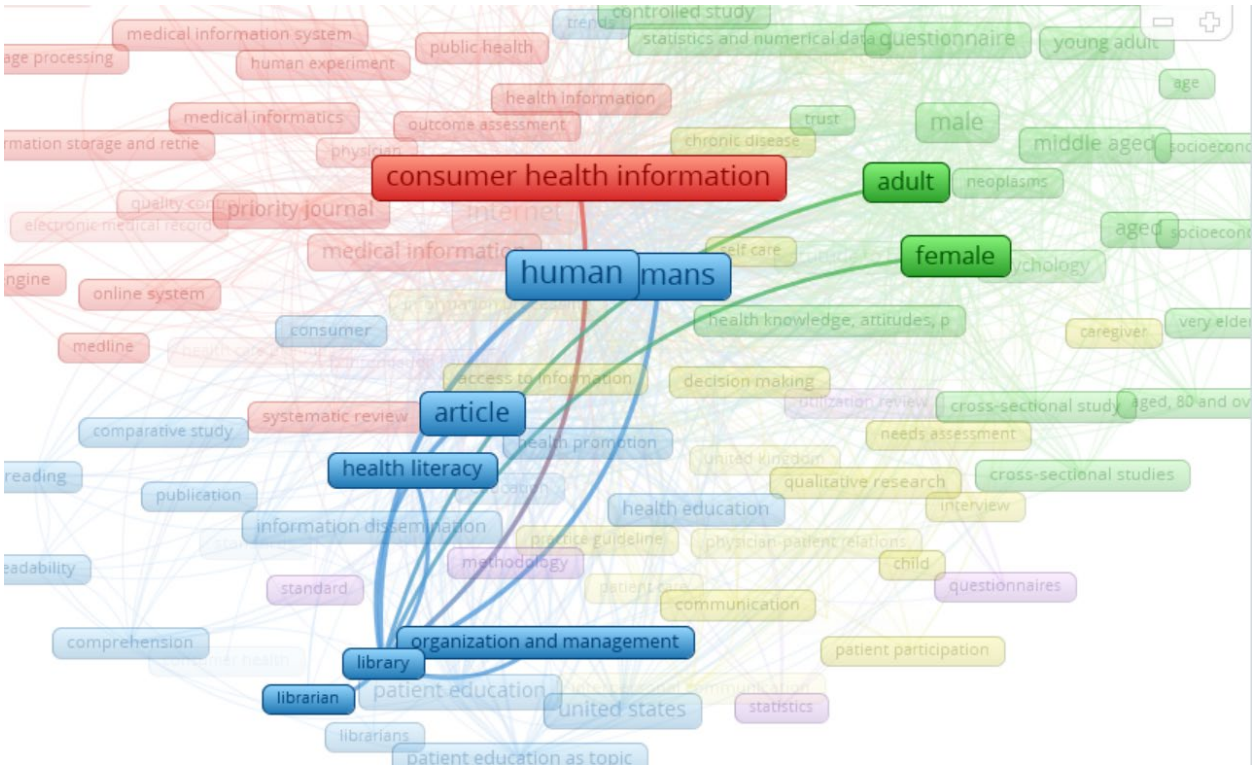
Cluster 3 is represented in blue and the main keywords are "human" and "humans" which both have 137 links. These keywords are strongly related to the terms article. health education, health literacy, information dissemination, library, organization and management, patient education, comprehension, and readability.



Cluster 4 is represented in yellow. The t most widely linked term is information processing. This cluster contains keywords like physician-patient relations, practice guidance, patient care, interpersonal communications, caregiver, chronic disease, self-care and social support. Cluster 5 is represented in purple. The main keyword in this cluster is methodology with 135 links, a total strength of 1145 and 89 occurrences. Other keywords include questionnaires, statistics, utilization review, and information service.



And finally, a visualization that illustrates the relationship between librarians, libraries and CHI.


controlled study
medical information system
public health
statistics and numerical data
questionnaire
young adult
human experiment
health information
age
medical informatics
outcome assessment
trust
male
chronic disease
middle aged
consumer health information
adult
neoplasms
priority journal
female
human
mans
online system
health knowledge, attitudes, p
consumer
caregiver
medline
access to information
decision making
article
systematic review
cross-sectional study
comparative study
needs assessment
health promotion
health literacy
qualitative research
cross-sectional studies
publication
health education
interview
information dissemination
methodology
child
standard
questionnaires
communication
comprehension
organization and management
patient participation
library
librarian
patient education
united states
statistics
librarians
patient education as topic

Works Cited

Consumer Health Informatics. (2018). MeSH Subject Scope Note. https://meshb.nlm.nih.gov/record/ui?ui=D048088

Demiris, G. (2016). Consumer Health Informatics: Past, Present, and Future of a Rapidly Evolving Domain. Yearbook of Medical Informatics, 25(S 01), S42–S47. https://doi.org/10.15265/IYS-2016-s005

Doll, T. (2019, March 11). LDA Topic Modeling. Medium. https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd

Joo, S. (2020, March 2). Topic Modeling [Lecture Notes]. LIS/ICT 662 - Data Analysis and Visualization - Spring 2020 - Week 8, University of Kentucky.

van Eck, N. J., & Waltman, L. (2014). Visualizing Bibliometric Networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), Measuring Scholarly Impact (pp. 285–320). Springer International Publishing. https://doi.org/10.1007/978-3-319-10377-8_13

Appendices

Appendix I

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
medic	inform	librari	patient	web	project	librarian	educ	health	studi	articl	associ	librari	develop	question
review	provid	public	inform	site	nation	hospit	program	consum	result	research	librari	servic	communiti	sourc
book	resourc	staff	center	onlin	outreach	role	scienc	inform	user	author	present	refer	mani	data
collect	health	train	famili	inform	medicin	profession	chang	literaci	evalu	issu	includ	discuss	scienc	search
guid	access	system	care	internet	nlm	support	profession	care	survey	practic	meet	technolog	peopl	content
resourc	need	state	univers	avail	institut	describ	student	initi	identifi	publish	mla	need	work	biomed
electron	find	increas	knowledg	websit	organ	respons	medic	address	conclus	trend	articl	report	social	languag
databas	locat	involv	materi	qualiti	paper	collabor	school	offer	object	journal	discuss	build	focus	general
list	seek	provis	clinic	select	local	academ	improv	patron	method	year	director	serv	creat	current
record	patron	event	medlineplus	includ	partnership	work	research	librari	assess	impact	member	digit	relev	specif
print	creat	awar	improv	medicin	plan	activ	continu	made	particip	valu	develop	manag	area	model
support	effect	focus	popul	consum	effort	play	cours	abil	conduct	time	american	space	challeng	medlin

Appendix II

Sample of the Topic Frequency within Documents

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	0.024345	0.080524	0.17603	0.041199	0.024345	0.029963	0.035581	0.052434	0.080524	0.097378	0.035581	0.035581	0.131086	0.06367	0.09176
2	0.07971	0.036232	0.047101	0.068841	0.112319	0.123188	0.036232	0.047101	0.123188	0.036232	0.047101	0.047101	0.068841	0.036232	0.09058
3	0.022222	0.05812	0.02735	0.052991	0.324786	0.022222	0.017094	0.02735	0.145299	0.145299	0.037607	0.02735	0.022222	0.022222	0.047863
4	0.026247	0.104987	0.081365	0.041995	0.057743	0.026247	0.034121	0.049869	0.175853	0.144357	0.041995	0.049869	0.081365	0.034121	0.049869
5	0.024155	0.089372	0.05314	0.045894	0.05314	0.031401	0.089372	0.154589	0.060386	0.205314	0.038647	0.031401	0.05314	0.024155	0.045894
6	0.022337	0.099656	0.063574	0.017182	0.073883	0.017182	0.058419	0.037801	0.063574	0.367698	0.017182	0.027491	0.053265	0.053265	0.027491
7	0.117048	0.086514	0.147583	0.124682	0.033079	0.025445	0.048346	0.033079	0.063613	0.048346	0.05598	0.033079	0.109415	0.033079	0.040712
8	0.057576	0.093939	0.075758	0.093939	0.057576	0.084848	0.057576	0.066667	0.130303	0.048485	0.048485	0.048485	0.057576	0.039394	0.039394
9	0.036415	0.095238	0.103641	0.053221	0.053221	0.053221	0.078431	0.070028	0.170868	0.061625	0.028011	0.028011	0.036415	0.086835	0.044818
10	0.029478	0.022676	0.199546	0.036281	0.056689	0.036281	0.077098	0.131519	0.15873	0.056689	0.036281	0.043084	0.036281	0.029478	0.049887

Appendix III

Main Information about data

Documents	1588
Sources (Journals, Books, etc.)	675
Keywords Plus (ID)	5573
Author's Keywords (DE)	3024
Period	2010 - 2020
Average citations per documents	12.06
Authors	4583
Author Appearances	5697
Authors of single-authored documents	233
Authors of multi-authored documents	4350
Single-authored documents	296
Documents per Author	0.346
Authors per Document	2.89
Co-Authors per Documents	3.59
Collaboration Index	3.37

Document types

ARTICLE	1162
BOOK	26
BOOK CHAPTER	60
CONFERENCE PAPER	157
EDITORIAL	9
LETTER	8
NOTE	10
REVIEW	154
SHORT SURVEY	2

Hit <Return> to see next table:
Annual Scientific Production

Year	Articles
2010	139
2011	138
2012	109
2013	173
2014	139
2015	177
2016	172
2017	160
2018	153
2019	167
2020	61

Annual Percentage Growth Rate -7.905965

Hit <Return> to see next table:
Most Productive Authors

	Authors	Articles	Authors	Articles Fractionalized
1	ZHANG Y	27	ZHANG Y	16.08
2	KESELMAN A	13	SMITH CA	8.00
3	SMITH CA	13	RUBENSTEIN EL	7.00
4	KIM S	11	YI YJ	6.50
5	OELSCHLEGEL S	11	CHARBONNEAU DH	5.50
6	PLUYE P	11	GRANT MJ	5.00
7	YI YJ	11	KIM S	4.96
8	DEMNER-FUSHMAN D	10	FLAHERTY MG	4.75
9	OH S	10	OH S	4.53
10	LI J	9	KESELMAN A	4.03

Hit <Return> to see next table:
Top manuscripts per citations

	Paper	TC	TCperYear
1	PIWEK L, 2016, PLOS MED	309	61.8
2	KONTOS E, 2014, J MED INTERNET RES	266	38.0
3	HEART T, 2013, INT J MED INFORMATICS	217	27.1
4	KUO AMH, 2011, J MED INTERNET RES	216	21.6
5	HU Y, 2010, COMMUN RES	202	18.4
6	SWAN M, 2012, J PERS MED	194	21.6
7	WANG Y, 2015, OBES REV	193	32.2
8	LAGAN BM, 2010, BIRTH	184	16.7

9	CHINN D, 2011, SOC SCI MED	170	17.0
10	ANKER AE, 2011, PATIENT EDUC COUNS	149	14.9

Hit <Return> to see next table:
Corresponding Author's Countries

	Country	Articles	Freq	SCP	MCP	MCP_Ratio
1	USA	570	0.5067	533	37	0.0649
2	CANADA	71	0.0631	59	12	0.1690
3	UNITED KINGDOM	64	0.0569	57	7	0.1094
4	AUSTRALIA	51	0.0453	43	8	0.1569
5	KOREA	37	0.0329	22	15	0.4054
6	GERMANY	28	0.0249	22	6	0.2143
7	NETHERLANDS	25	0.0222	21	4	0.1600
8	ITALY	19	0.0169	16	3	0.1579
9	CHINA	16	0.0142	6	10	0.6250
10	IRAN	15	0.0133	12	3	0.2000

SCP: Single Country Publications

MCP: Multiple Country Publications

Hit <Return> to see next table:
Total Citations per Country

	Country	Total Citations	Average Article Citations
1	USA	8098	14.21
2	UNITED KINGDOM	1026	16.03
3	NETHERLANDS	972	38.88
4	AUSTRALIA	885	17.35
5	CANADA	789	11.11
6	GERMANY	333	11.89
7	ISRAEL	249	41.50
8	KOREA	216	5.84
9	IRELAND	202	33.67
10	NORWAY	193	16.08

Hit <Return> to see next table:
Most Relevant Sources

	Sources	Articles
1	JOURNAL OF MEDICAL INTERNET RESEARCH	73

2	JOURNAL OF CONSUMER HEALTH ON THE INTERNET	60
3	HEALTH INFORMATION AND LIBRARIES JOURNAL	57
4	PLOS ONE	40
5	JOURNAL OF HEALTH COMMUNICATION	37
6	JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION	36
7	STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS	32
8	JOURNAL OF HOSPITAL LIBRARIANSHIP	31
9	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	26
10	JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY	21

Hit <Return> to see next table:

Most Relevant Keywords

	Author Keywords (DE)	Articles	Keywords-Plus (ID)	Articles
1	CONSUMER HEALTH INFORMATION	196	CONSUMER HEALTH INFORMATION	977
2	INTERNET	179	HUMAN	890
3	HEALTH LITERACY	130	INTERNET	862
4	HEALTH INFORMATION	74	HUMANS	762
5	PATIENT EDUCATION	58	FEMALE	699
6	SOCIAL MEDIA	57	MALE	620
7	CONSUMER HEALTH	41	ADULT	580
8	READABILITY	39	ARTICLE	519
9	INFORMATION SEEKING	37	MIDDLE AGED	401
10	HEALTH EDUCATION	34	AGED	328