Journal of Library and Information Studies 18:1 (June 2020) pp.1-23
doi: 10.6182/jlis.202006_18(1).001

Exploring the Domain of Information "Users": Semantic Analysis of Wikipedia Articles

Soohyung Joo[1]

Abstract

This study aims to explore the domain of information users based on the textual analysis of Wikipedia. The study collected 132 Wikipedia articles related to information users. Then, multiple existing methods of textual analysis were applied to explore prevalent concepts and topics in relation to information users. The text analysis methods applied in this study include: Term frequency analysis, the Latent Dirichlet Allocation topic models, hierarchical clustering, dimensional reduction, and term co-occurrence analysis. The findings reveal that Wikipedia covers various topics of the domain of information users. The domain of information users involves four fundamental concepts, including "users," "behaviors," "systems/technology," and "information objects." In addition, the study identifies different areas of the domain, ranging from information search behavior, information retrieval, human-computer interaction, user experience, human factors, and to others.

Keywords: Information Users; Domain Analysis; Textual Analysis; Wikipedia; User Studies

1. Introduction

Users have been considered one of the key concepts of interest in information sciences. Users create and distribute knowledge and information, and then, other users may search, access, and use those information to fulfill their own information needs in various contexts. In online environments, users have to interact with systems to store, share, or retrieve information according to their needs and tasks. All those information and systems ultimately exist to satisfy users' certain information needs. In this way, users are placed in the center of an information eco-system. There is a plethora of literature that studied "users" in information sciences as well as in other disciplines, such as computer sciences and communications. Thus, it is meaningful to explore what topics or concepts have been discussed in relation to information users to help the research community have a better understanding of the domain of information users.

Traditional bibliographic analysis of scholarly articles has been widely used to survey the domain of user research in information sciences (e.g., Lee & Kam, 2018; Vošner, Kokol, Bobek, Železnik, & Završnik, 2016). However, the body of literature regarding information users is extremely vast, so it might not easy to collect and investigate those heterogeneous literature comprehensively that might include a tremendous amount of noise and irrelevant documents. Wikipedia can be an alternative source of data in that it includes a succinct, well-structured format of knowledge about certain domains. Wikipedia refers to an

[1] School of Information Science, University of Kentucky, USA
E-mail: soohyung.joo@uky.edu

臺灣大學學術期刊資料庫

online encyclopedia and its content is collectively created and edited by volunteer users around the world (Wikipedia, 2019). Wikipedia is made of a collaborative effort to represent human knowledge in a structured way, and more importantly, it is open to the public (Jatowt, Kawai, & Tanaka, 2016). The collective contributions from domain expert editors across all over the world have made Wikipedia to serve as a huge, constantly evolving knowledge source that covers numerous topics (Halavais & Lackaff, 2008; Iba, Nemoto, Peters, & Gloor, 2010; Milne & Witten, 2013). Regardless of the debates on the quality of Wikipedia (Mühlhauser & Oser, 2008), it has been considered a rich source of knowledge that describes knowledge of certain domains in an organized format. In addition, Wikipedia content has been reported as it keeps a sufficient level of accuracy in general (Chesney, 2006; Giles, 2005; Wilkinson & Huberman, 2007).

This study uses Wikipedia as a source to explore the domain of information users. Wikipedia stores and updates a large set of articles concerning information user related topics, and the analysis of those articles makes us to better understand what topics are prevailing in relation to information users. This study intends to explore various topics and concepts within the domain of information users based on the content analysis of Wikipedia articles. We employed multiple existent textual analysis methods, including term frequency analysis, probabilistic topic models, clustering and dimension reduction, and term co-occurrence analysis.

2. Literature Review

2.1 User research areas in information sciences

Users have been one of the main components in an information ecosystem, and the information science and other related research disciplines have studied users with regards to information access, retrieval, use, technologies, ethics, usability, and system design. Particularly, many studies in information science have investigated behavior, interaction, engagement, and relationships between users and information sources or users and systems. In this literature review section, it is not possible to review all variety of relevant concepts due to the space limitation. Instead, several major concepts in user research are briefly reviewed.

Skov (2009) identified different levels of relationships between the user and the information. According to Skov (2009), the concepts of information behavior, information seeking, and information retrieval are hierarchically nested. Information behavior is a broad concept at the highest level, and it covers activities a person may engage in when identifying his/her needs of information, searching for such information and using the information (Wilson, 1999, 2000). Information seeking is a narrower concept of information behavior and is considered a person's purposive seeking activities for information to satisfy an underlying goal (Wilson, 2000). Particularly, Wilson's research (Wilson, 1981) on user studies and information needs made a significant contribution to the development of information sciences as it discussed fundamental issues of the nature of information, information need, and information seeking and information behavior (Bawden, 2006).

臺灣大學學術期刊資料庫

Information retrieval concerns formal algorithmic processes of document representation, storage, searching, and finding of potential information sources perceived relevant to a user's need (Ingwersen & Järvelin, 2006). Traditionally, information retrieval has involved the development of indexing techniques, query analysis, and effective matching algorithms. Then, the two fundamental measures, including precision and recall, have been widely used to assess the outcomes of retrieval results (Harman & Voorhees, 2006). The other stream of information retrieval research has extensively focused on users. Researchers in information retrieval have paid attention to users' interaction with information retrieval systems (Callan et al., 2007). Human-computer information retrieval, also called interactive information retrieval, is the research area where it examines interactive communication process occurring during retrieval of information (Ingwersen & Järvelin, 2006). Borlund (2000, 2003) suggested task-based evaluation methods for interactive information retrieval by incorporating user perspectives on relevance, which dealt with multidimensional and dynamic aspects of relevance. Some researchers have focused on users' engagement with information retrieval and information search (O'Brien & Toms, 2008; Xie, 2008; Xie, Joo, & Bennett-Kapusniak,, 2017). Xie (2008) defined user engagement as a user's behavioral and cognitive intention to change the information retrieval process affecting the retrieval outcomes. Additionally, in information sciences, there have been efforts to develop models to explain complex patterns of user behavior, such as Wilson's model of information behavior (Wilson, 1999), Kuhlthau's information seeking process model (Kuhlthau, 1991, 1993) and Vakkari's model of information search process (Vakkari, 2000a, 2000b, 2001).

User experience has emerged as an important area in both academia and industries. User experience is originally a broad concept that covers various aspects of users' interaction with products, such as the way a user feels, how well a user understands, and how well the product serves its purposes (Alben, 1996). In the information science, user experience is considered a multi-dimensional concept that can be used to understand different aspects of interactions between the users and information systems (O'Brien & Toms, 2013; Xu, 2015). User experience design is a widespread paradigm of system design practice that that emphasizes users in the process of design process (Zhao, Liu, Tang, & Zhu, 2013). There are various approaches and methods for user experience design, such as persona analysis, task analysis, user stories, wireframing, and prototyping (Da Silva, Silveira, Maurer, & Hellmann, 2012; Nielsen, Hansen, Stage, & Billestrup, 2015; Rohrer, 2014). Then, usability is closely related to user experience. Usability concerns the question of whether a system is good enough to satisfy users and it could help users do their jobs better (Morse, 2002; Nielsen, 1993). In general, usability consists of multiple attributes, such as effectiveness, efficiency, learnability, satisfaction, and others (International Organization for Standardization, 1998; Jeng 2005; Joo & Lee, 2011; Nielsen, 1993). There are a range of specific methods and techniques to assess usability, such as experimental design, heuristic evaluation, cognitive walkthrough, and others (Hollingsed & Novick, 2007; Nielsen, 1994).

臺灣大學學術期刊資料庫

2.2 Analysis of Wikipedia documents

Wikipedia has been an ample source of knowledge created collectively by domain experts for certain areas, and researchers have explored the content of Wikipedia articles using various textual analysis techniques. It has been considered a goldmine of knowledge for a large number of research communities (Medelyan, Milne, Legg, & Witten, 2009). Ni, Sun, Hu, and Chen (2009) employed a topic model algorithm to identify multilingual topics from a set of large scale and multilingual Wikipedia articles. They proposed and tested a novel approach, called the multilingual topics, to specifically analyze multilingual documents from Wikipedia. Halavais and Lackaff (2008) were interested in the coverage of Wikipedia across different disciplines. Their study was conducted based on the assumption that some areas could be more covered than others in Wikipedia. They found that topical coverage would be dependent on the interests of its contributors in Wikipedia, and some topical areas exhibited more coverage with rapid expansion, such as some popular cultural topics or physical science. Iba et al. (2010) analyzed editing behaviors of Wikipedia content contributors based on social network analysis and identified most creative Wikipedia contributors, particularly those who initiated new high-quality articles. They identified two types of Wikipedia editors based on the analysis of editing patters, so called the mediators and the zealots.

Heilman and West (2015) tried to measure the quality of the production and consumption of medical content in Wikipedia using multiple approaches. They first quantified the amount of medical content based on the analysis of document size and citations. They also investigated the quantity and characteristics of Wikipedia's medical contributors, including the contributors' participation trends and demographic information. Jatowt et al. (2016) explored historical data collected from Wikipedia in an effort to investigate its time-related characteristics, focusing on historical persons. They investigated the characteristics of historical information about the chosen historical figures, and attempted to quantify users' interest in those documents based on the analysis of different cues, such as hyperlink structure and article view logs. Mehdi, Okoli, Mesgari, Nielsen, and Lanamäki (2017) systematically reviewed academic literature that analyzed Wikipedia documents. They found that Wikipedia has been one of the popular research topics in the computer science discipline, particularly in relation to information retrieval, natural language processing, and ontology. They claimed that Wikipedia provided rich, large text content in certain domains and it motivated many computer science researchers to explore different aspects of Wikipedia using textual analysis methods. Recently, Wang (2018) investigated the content of family health topics, such as child maltreatment, family planning, and women's health, on Wikipedia. She applied text mining, especially the self-organizing map technique, to explore the characteristics and popularity of those selected health topics.

Credibility and accuracy of Wikipedia content has also been examined in prior studies. Kräenbring et al. (2014) examined the accuracy of drug information in the German and English versions of Wikipedia. They compared the Wikipedia content on the topic with educational

臺灣大學學術期刊資料庫

pharmacology textbooks. Their study revealed that Wikipedia could be considered an accurate and comprehensive source of drug information for both German and English documents. They claimed that the content from Wikipedia would be acceptable for the undergraduate level of medical education. Similarly, Azer (2014) examined the credibility of Wikipedia sources in two medical areas, including gastroenterology and hepatology. He selected 39 articles in the chosen areas, and then, critically evaluated the quality of the information. The investigation revealed that Wikipedia covered a wide range of topics in the areas of gastroenterology and hepatology, and the documents were well organized by following the standardized structure of Wikipedia. However, he pointed out several problems of Wikipedia articles, such as a limited depth of content and insufficient references and citations.

In addition, there have been efforts to understand the organization, clustering, or networks of documents in Wikipedia. Kittur, Chi, and Suh (2009) proposed a novel method for mapping Wikipedia articles based on the distribution of different topics in Wikipedia and further analyzed the patterns of how those topics changed over time. They mapped the topical distribution in Wikipedia into several categories, such as "culture and the arts," "people and self," "geography and places," "society and social sciences," and others. Wang, Joo, and Lu (2014) explored topical themes in the field of data science based on the clustering analysis of Wikipedia articles. They collected 550 documents related to data science research, and conducted hierarchical clustering and principal component analysis to identify the structure of knowledge in the field of data science reflected in Wikipedia. Chen, Lin, and Evans (2019) investigated the network of mathematicians based on the analysis of hyperlinks in Wikipedia articles. Using social network analysis and following visualization, they identified Hilbert and Newton as the most important mathematicians.

The following research questions guide the investigation of this study:

- What are the prevailing terms and topics in the domain of information users in Wikipedia?
- Which clusters and dimensions are identified from the Wikipedia documents related to information users?
- What are the key concepts representing the domain of information users based on the analysis of Wikipedia?

3. Methods

The purpose of this study is to explore key concepts and topics in the information users domain through the analysis of Wikipedia articles. To collect relevant articles, we first chose four seed topics, including "information seeking," "human computer interaction", "user experience", and "information retrieval." Then, we expanded them by collecting relevant documents based on the following methods: First, we collected documents linked in the "See Also" section in each seed document. Second, relevant documents were chosen from the top fifty results returned from the keyword search of each seed word. We filtered out irrelevant or too short documents using the following criteria: any document less than 300 words; documents about commercial products/software (e.g., Android Marshmallow),

臺灣大學學術期刊資料庫

totally non-relevant documents (e.g., The Seeker - Australian pop band); the names of organizations/ groups (e.g., Center for Intelligent Information Retrieval); and the names of publication venues (e.g., ACM Transactions on Computer-Human Interaction). In addition, any document about a certain person were not included in the analysis. After the screening process, we ended up with 132 relevant documents that could represent the domain of information users.

All collected documents underwent the text pre-processing including tokenization, stopword removal, and stemming. Then, we employed existing textual analysis methods to answer the research questions. First, we tallied term frequency to explore most frequent terms in the corpus. Term frequency analysis was conducted at the fundamental level to find out prominent terms in the domain. Second, the latent Dirichlet allocation (LDA) topic model was used. The LDA topic model is an unsupervised machine learning technique that was developed to discover or detect hidden topics/themes based on the analysis of the semantic structure of the document collection (Blei, 2012; Blei, Ng, & Jordan, 2003). LDA topic modeling was used to detect prevailing, topical terms that made up primary topics in the domain of information users. We tested different sets of topics in the topic model analysis, including 12, 15, 18, 20, and 25 topics respectively, and then, we determined that the model of 20 topics would generate most coherent topics to be interpreted. Articles were assigned to topics based on calculated topic probabilities. Third, hierarchical clustering was conducted at the document level. Hierarchical clustering is a technique to explore a hierarchical structure of clusters that could exhibit mutually exclusive subsets (Everitt, Landau, & Leese, 2001; Maimon & Rokach, 2005). To be more specific, a dendrogram was generated based on the Ward method to analyze the relationships among the documents. Fourth, principal component analysis (PCA) complemented the clustering analysis to examine different dimensions of the collected documents. It was used to uncover dimensionality of observations (Jackson, 2005). PCA is considered an effective method to identify the structure of web documents in a certain domain (Kipp, Joo, & Choi, 2013). Finally, we analyzed co-occurrences of observed terms and created a visual map highlighting the relationships among the key terms using VOSviewer (http://www.vosviewer.com). It is a software tool to visualize the relationships based on network data, and it allows to extract keyword terms and analyze their co-occurrences in a network space. The tool provides a text mining functionality to identify phrases from large amounts of unstructured text and visualizes the mapping of those terms based on co-occurrences (Van Eck & Waltman, 2011, 2019). This study utilized such text mining functionality in the VOSviewer tool to identify prominent terms in the co-occurrence map.

4. Results

First, we conducted term frequency analysis. After removing stop words, we observed unique 15,104 terms, consisting of 177,094 tokens. The first ranked term was “inform,” which made up 2.20% of the total tokens, followed by “user (1.21%).” As “information” and “users” are the two main concepts in user related documents in

臺灣大學學術期刊資料庫

information sciences, those two terms were most frequently observed. Then, "design (1.09%)," "comput (1.00%)," and "system (0.96%)" were ranked third, fourth, and fifth respectively. The term "retriev" was ranked 6th, which reflects articles related to information retrieval. Also, we found that "interact (0.67%)" and "human (0.64%)" were ranked amongst top ten at 7th and 8th respectively. These two terms might be associated with the area of human-computer interaction. We also observed that some terms would were associated with information behavior models and theories, for example, "model (0.40%)," "need (0.34%)," "theori (0.33%)," "behavior (0.30%)," and "seek (0.19)." In addition, we found that some frequent terms refer to research or academia, such as "research (0.49%)," "scienc (0.45%)," "studi (0.30%)," and "journal (0.22%)."

As shown in Table 2, we extracted twenty topics from the document corpus using the LDA topic mode. Topic 2, Topic 11, and Topic 15 are related to information retrieval, document indexing, and information retrieval measurement. Topic 19 represents user experience design and usability. Topic 14 reveals the topic of human computer interaction. Topic 13 is relevant to the library and information science. Topic 6 concerns web and graphic design. Topic 1, Topic 10, and Topic 17 represent the area of information seeking and behavior and associated theories and models. In this way, we observed that various aspects of user related topics have been discussed in the Wikipedia environment.

Next, we conducted a hierarchical clustering analysis in an attempt to identify the clusters of related documents. For clustering analysis, we removed less meaningful terms. We calculated the importance of terms using the *tf-idf* values, and then, chose 4,512 terms to be included in the analysis. We used this refined dataset for hierarchical clustering as well as dimensional reduction analysis. Using the refined data, we conducted hierarchical clustering based on the Ward method. From the analysis, we identified six larger clusters and twelve sub-clusters as shown in Table 3. The higher level clusters include: "information retrieval and search," "user experience," "information seeking and retrieval," "research disciplines," "information retrieval evaluation," and "human factors and theories." Then, more specific 12 sub-clusters were defined, such as "information searching," "user research and interface," "user experience and inferface," "human centered design," and others. Many of the collected documents were related to each other in some ways as they were chosen from the common theme of information users. Thus, we found that clusters are somewhat overlapped with each other.

We also conducted principal component analysis (PCA) to explore the different dimensions of the collected document set. As suggested from the hierarchical clustering analysis, we assumed that twelve components would well portrait the entire domain of the information user domain. Those twelve components explained approximately 51.4% of the total variance at the eigenvalue of 1.351. As shown in Figure 1, the scree plot also indicates that a long tail starts from around the twelfth component.

The PCA results revealed distinct latent dimensions of the information user domain. Table 4 presents the extracted components and associated representative documents for each

臺灣大學學術期刊資料庫

Table 1. Most Frequently Observed Terms (Top 100 Ranked)

Rank	Term	Frequency	Percent	Rank	Term	Frequency	Percent
1	inform	3,904	2.20	52	acm	392	0.22
2	user	2,137	1.21	53	usabl	384	0.22
3	design	1,926	1.09	53	journal	384	0.22
4	comput	1,766	1.00	55	societi	383	0.22
5	system	1,702	0.96	55	librari	383	0.22
6	retriev	1,461	0.82	57	mathrm	377	0.21
7	interact	1,190	0.67	58	person	375	0.21
8	human	1,133	0.64	58	ontolog	375	0.21
9	research	865	0.49	60	refer	367	0.21
10	data	841	0.47	61	relev	366	0.21
11	scienc	791	0.45	62	result	350	0.20
12	search	768	0.43	63	softwar	345	0.19
13	interfac	741	0.42	63	provid	345	0.19
14	document	739	0.42	65	queri	345	0.19
15	develop	725	0.41	66	proceed	340	0.19
16	process	720	0.41	67	perform	337	0.19
17	experi	717	0.40	67	seek	337	0.19
18	model	709	0.40	69	learn	334	0.19
19	web	665	0.38	69	media	334	0.19
20	manag	664	0.37	71	task	333	0.19
21	need	601	0.34	72	custom	332	0.19
22	theori	588	0.33	73	evalu	327	0.18
23	term	576	0.33	74	field	323	0.18
24	technolog	570	0.32	75	mine	321	0.18
25	social	562	0.32	76	specif	317	0.18
26	base	556	0.31	77	imag	316	0.18
27	communic	553	0.31	78	program	314	0.18
28	time	547	0.31	79	displaystyl	311	0.18
29	behavior	540	0.30	80	analysi	304	0.17
30	studi	530	0.30	80	mani	304	0.17
31	includ	522	0.29	82	text	303	0.17
32	work	492	0.28	83	origin	302	0.17
33	knowledg	490	0.28	84	group	297	0.17
34	relat	487	0.27	85	control	296	0.17
35	applic	485	0.27	85	understand	296	0.17

(continued)

臺灣大學學術期刊資料庫

Table 1. Most Frequently Observed Terms (Top 100 Ranked) (continued)

Rank	Term	Frequency	Percent	Rank	Term	Frequency	Percent
36	method	476	0.27	87	effect	294	0.17
37	confer	475	0.27	88	factor	292	0.16
38	test	465	0.26	89	level	289	0.16
39	product	451	0.25	90	creat	288	0.16
40	content	437	0.25	91	associ	286	0.16
41	exampl	419	0.24	92	problem	284	0.16
41	intern	419	0.24	93	approach	277	0.16
41	organ	419	0.24	93	structur	277	0.16
44	engin	413	0.23	95	univers	276	0.16
45	differ	412	0.23	95	busi	276	0.16
46	languag	410	0.23	97	oper	275	0.16
47	isbn	403	0.23	97	support	275	0.16
48	machin	401	0.23	99	individu	271	0.15
49	activ	397	0.22	100	requir	270	0.15
49	peopl	397	0.22	100	word	270	0.15
51	index	393	0.22	…	…	…	…

component. Twelve dimensions extracted include: D1. Information seeking and searching, D2. User interface, D3. User experience design, D4. Information retrieval, D5. Human factors, D6. Human computer interaction, D7. Information retrieval evaluation and models, D8. Indexing and organization, D9. Image retrieval, D10. Web and data mining, D11. Customer experience, and D2. Theories. We found that several documents were associated more than one component. For example, D2 (User interface) and D3 (User experience design) were somewhat overlapped. The document of "User_interface_design" showed a loading of .702 for D2 and that of .502 for D3. Similarly, the document of "Usability" exhibited the factor loadings of .580 for D2 (User interface) and .492 for D3 (User experience design) respectively. Also, there were a couple of documents associated both D3 (User experience design) and D5 (Human factors), such as "Human_centered_design" and "Human_factors_and_ergonomics." In addition, we found that D5 (Human factors) and D6 (Human computer interaction) were moderately related to each other. The documents of "Human_computer_interaction," "Human_centered_computing" and "Interactivity" showed more than a loading of 0.25 in both the components (D5 and D6). Some distinct topics in information retrieval were grouped as separate components, such as information retrieval evaluation and models (D7), indexing and organization (D8), and image retrieval (D9). Also, we observed a distinct dimension for the theory related documents (D12).

臺灣大學學術期刊資料庫

Table 2. LDA Topic Modeling Results (20 Topics)

Topic 1 (3.0%)	**Topic 2 (12.9%)**	**Topic 3 (3.0%)**	**Topic 4 (2.3%)**	**Topic 5 (1.5%)**
behavior	retriev	media	custom	ontolog
theori	search	social	experi	literaci
intent	inform	robot	social	educ
control	imag	gratif	manag	inform
predict	queri	research	busi	student
Self	collabor	human	employe	retriev
belief	system	theori	organ	librari
perceiv	relev	network	organiz	archiv
health	base	communic	compani	nation
action	user	audienc	job	learn
Topic 6 (4.5%)	**Topic 7 (0.8%)**	**Topic 8 (3.0%)**	**Topic 9 (3.8%)**	**Topic 10 (3.0%)**
design	test	data	comput	active
web	ture	mine	system	theory
graphic	intellig	societi	memori	person
content	machin	web	data	music
websit	human	knowledg	program	manag
page	comput	structur	pick	pim
respons	artifici	machin	database	cognit
creat	game	network	oper	brows
monitor	question	pattern	softwar	psychology
layout	mind	learn	scienc	tool
Topic 11 (6.8%)	**Topic 12 (3.8%)**	**Topic 13 (9.1%)**	**Topic 14 (9.1%)**	**Topic 15 (6.8%)**
index	time	inform	interact	mathrm
document	movement	scienc	interfac	displaystyl
languag	mind	system	comput	document
term	model	librari	user	precis
word	oper	technolog	human	recal
search	human	societi	system	retriev
model	bodi	organ	design	relev
text	interact	studi	human	frac
retriev	level	journal	devic	term
extract	doi	knowledg	machin	posit
Topic 16 (9.8%)	**Topic 17 (3.8%)**	**Topic 18 (0.8%)**	**Topic 19 (11.4%)**	**Topic 20 (0.8%)**
develop	uncertainti	acm	user	human
process	inform	confer	design	ergonom
need	seek	proceed	usabl	Factor
research	communic	intern	experi	stori
includ	theori	web	interact	engine
work	interact	inform	product	work
relat	behavior	search	system	method
system	reduct	tempor	method	psycholog
exampl	individu	retriev	test	incid
comput	reduc	sigir	interfac	industri

臺灣大學學術期刊資料庫

Table 3. The Clustering Analysis Results

Clusters	Sub-clusters	Representative documents (title)
Information searching and user research	C1. Information searching	[25 documents] Dwell_time_information_retrieval, Information_retrieval_query_language, Information_access, Information_scientist, Information_retrieval_applications, Faceted_search, Information_foraging, Browsing, Serendipity, Thesaurus_information_retrieval, etc.
	C2. User research and interface	[17 documents] User_research, Experience_architecture, Real_user_monitoring, Human_interface_guidelines, Human_machine_system, Sensemaking_information_science, Natural_mapping_interface_design, Cognitive_walkthrough, Critical_incident_technique, etc.
User experience	C3. User experience and interface	[11 documents] User_experience, User_interface_design, Graphical_user_interface, Heuristic_evaluation, User_computing, First_time_user_experience, End_user, User_story, etc.
	C4. Human centered design	[17 documents] Human_centered_design, Human_centered_computing, Interactivity, Interaction_technique, Mobile_interaction, Information_design, Information_architecture, Interaction_design_pattern, etc.
Information seeking and retrieval	C5. Image retrieval	[3 documents] Image_retrieval, Reverse_image_retrieval, Content_based_image_retrieval
	C6. Information seeking and needs	[6 documents] Information_seeking_behavior, Information_seeking, Information_needs, Comprehensive_model_of_information_seeking, Wilsons_model_of_information_behavior, Information needs
	C7. Information retrieval	[15 documents] Information_retrieval, Legal_information_retrieval, Relevance_feedback, Language_model, Music_information_retrieval, Human_computer_information_retrieval, Okapi_BM25, Relevance_(Information_retrieval), Knowledge_retrieval, etc.
	C8. Indexing and others	[7 documents] Search_engine_indexing, Subject_indexing, tf_idf, etc.
Research disciplines	C9. Research disciplines and areas	[6 documents] Information_science, Information_system Library_and_information_science, Information_management, Personal_information_management, Information_society
Information retrieval evaluation	C10. Information retrieval evaluation	[2 documents] Evaluation_measures_(information_retrieval), Precision_and_recall
Human factors and theories	C11. Human factors	[2 documents] Human_factors_and_ergonomics, Human_robot_interaction
	C.12 Information theories	[5 documents] Activity_theory, Collaborative_information_seeking, Theory_of_planned_behavior, Uses_and_gratification_theory, Onboarding
Others (not categorized the clusters above)		Information literacy, Temporal_information_retrieval, Turing_test, Uncertainty_reduction_theory, Customer_experience, Data_mining, Computer_science, Ontology_(information_science)

臺灣大學圖書館期刊資料庫

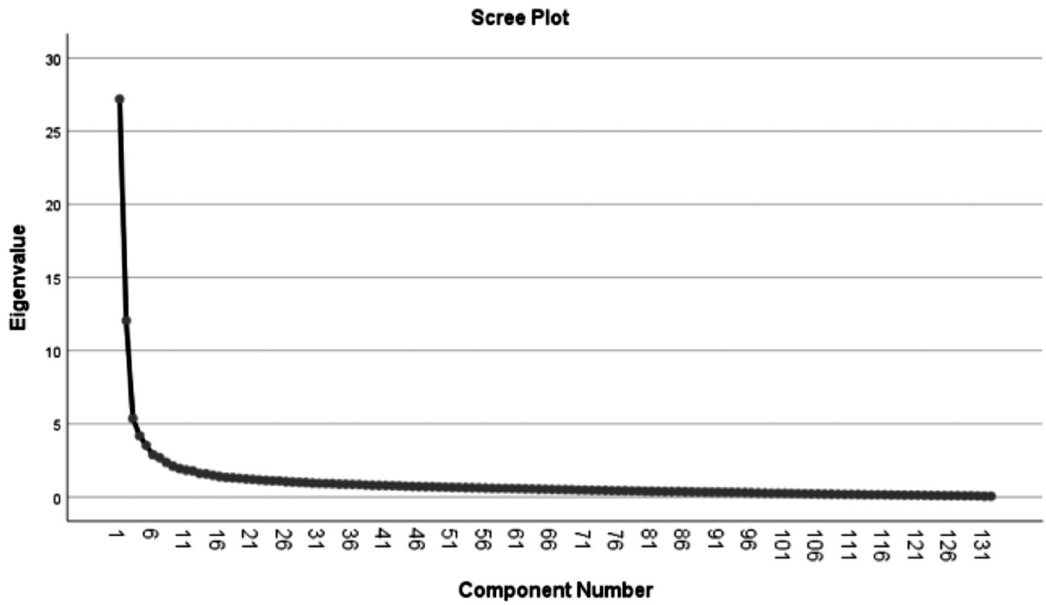


Figure 1. Scree Plot

Finally, we conducted the term co-occurrence analysis and visualized the results. As shown in Figure 2, "user," "information," and "system" are the three most prominent terms in the visualization. That is, we can infer that these three concepts construct the basis of the information user domain in information sciences. Then, the visualization indicates that "design," "interface," "interaction," "theory," "behavior," "research," "document," "use" and "person" were among the secondary noticeable terms.

5. Discussion and Conclusion

This study attempted to explore one of the fundamental components in information sciences, "users," and explored related concepts based on the analysis of Wikipedia articles. We found that Wikipedia covers various aspects of user related topics, and the analysis of those documents helped understand the entire sphere of user related topics. We retrieved and collected various articles related to users from Wikipedia, and then, adopted text mining techniques to understand what topics and concepts have been discussed collectively.

Most frequent terms highlighted that "information (1st ranked)" and "user (2nd ranked)" are the two fundamental concepts in the information users domain. In an information eco-system, users are considered a main entity who creates, organizes, stores, shares, searches, and consumes information. The domain of information users constructs various topics and relationships surrounding those two terms. Then, "design", "compute" and "system" made up another frequently observed terms from the collected document set, ranked 3rd, 4th, and 5th respectively. Computers and systems are

臺灣大學學術期刊資料庫

Table 4. The Principal Component Analysis Results

Dimension	Document
D1. Information seeking and searching	[29 documents] Information_science, Information_seeking_behavior, Information_seeking, Information_needs, Information_management, Library_and_information_science, Information_literacy, etc.
D2. User interface	[20 documents] User_interface , User_(computing), End_user, User_interface_design, User_Advocacy, User_experience, Heuristic_evaluation, User_story, Graphical_user_interface, Usability, Human_interface_guidelines, etc.
D3. User experience design	[13 documents] Graphic_design, User_experience_design, Interaction_design, Web_design, User_centered_design, Website_wireframe, Persona_(user_experience), etc.
D4. Information retrieval	[16 documents] Information_retrieval, Information_retrieval_applications, XML_retrieval, Cognitive_models_of_information_retrieval, Cross_language_information_retrieval, Knowledge_retrieval Information_retrieval_query_language, Document_retrieval, etc.
D5. Human factors	[14 documents] Human_factors_and_ergonomics, Human_processor_model, Human_centered_design, Human_machine_system, Human_robot_interaction, Keystroke_level_model, etc.
D6. Human computer interaction	[8 documents] Human_computer_interaction, Interactivity, Computer_science, Human_centered_computing, Mobile_interaction Modality_(human_computer_interaction), etc.
D7. Information retrieval evaluation and models	[8 documents] Evaluation_measures_(information_retrieval), Precision_and_recall, Okapi_BM25, tf_idf, Boolean_model_of_information_retrieval, Relevance_feedback, etc.
D8. Indexing and organization	[5 documents] Subject indexing, Controlled_vocabulary, Search_engine_indexing, Thesaurus_(information_retrieval), Knowledge_organization
D9. Image retrieval	[4 documents] Reverse_image_search, Image_retrieval, Content-based image retrieval, Faceted_search
D10. Web and data mining	[6 documents] Web_mining, Data_mining, Information_extraction, Ontology_(information_science), etc.
D11. Customer experience	[3 documents] Customer_experience, Quality_of_experience, Chief_experience_officer
D12. Theories	[5 documents] Uses_and_gratifications_theory, Uncertainty_reduction_theory, Theory_of_planned_behavior, etc.

the necessary tool to organize, store, and search information, and the design of those systems for users has been of interest in the information users domain. We also found that "retriev (6th)" and "interact (7th)" are also among the top ranked terms. Information retrieval has been long a major research agenda in both library and information sciences and computer sciences.

臺灣大學學術期刊資料庫

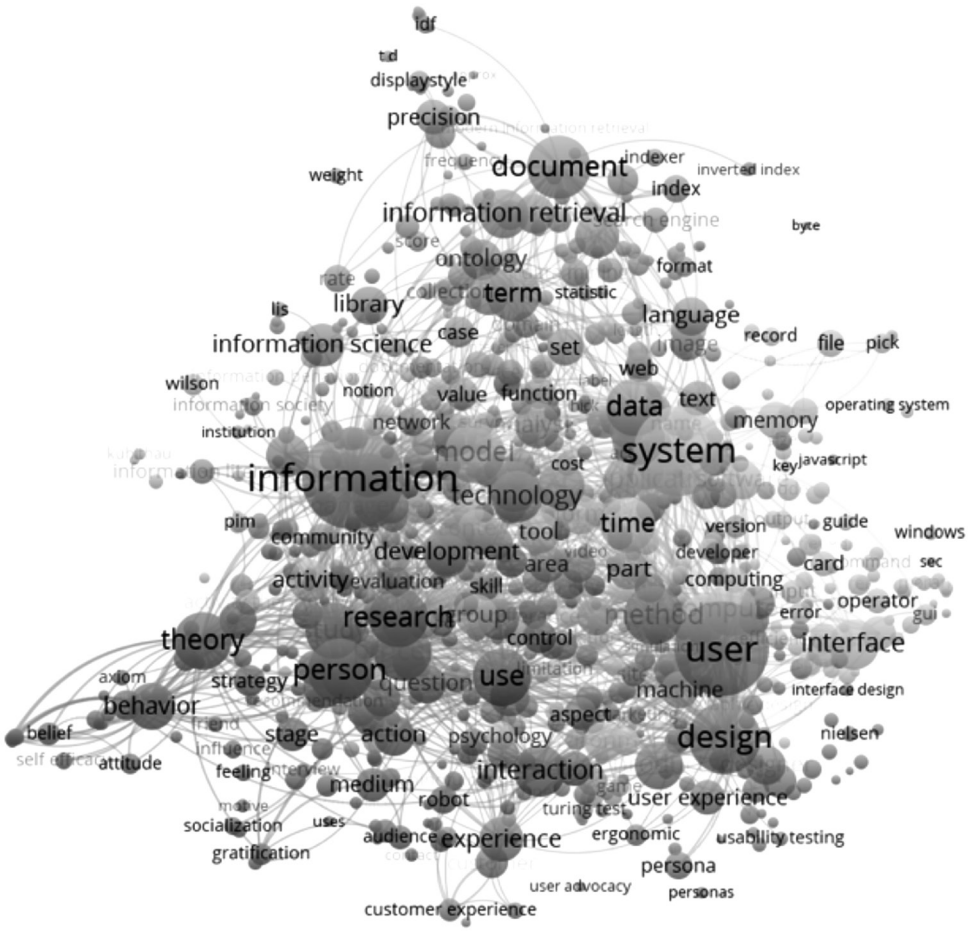


Figure 2. Term Co-occurrence Analysis: A Term Link Map

"Interaction" is another important keyword in user involved studies in information sciences and other disciplines, particularly, human-computer interaction emergied as a distinct research area. Different terms representing users were also observed among top frequent terms, such as "human (8th)," "peopl (49th)," "person (58th)," "group (84th)" and "individu (99th)." Then, some terms indicate the behavioral side of users, such as "interact (7th)," "search (12th)," "need (21th)," "behavior (29th)," "queri (65th), " "seek (67th)," and "learn (69th)." There are also various terms that indicate related systems and technologies, for example, "system (5th)," "interfac (13th)," "web (19th)," "technolog (24th)," and "softwar (63th)." Finally, we could see several terms that represent information objects, such as "inform (1st)," "data (10th)," "document (14th)," and "knowledge (33th)." In this way, most frequent terms explicitly reveal the distinct four core realms of components,

臺灣大學學術期刊資料庫

including users, behaviors, technologies, and information objects, in the domain of information users. The relationships among these four core realms can be further explained by models and theories. That is, we can infer that the terms of "model (18th)" and "theori (22nd) " are used to explain the relationships among the core realms. As shown in Figure 3, the four core realms extracted from the observation of the frequent terms can construct the domain of information users.

The LDA topic models explicitly showed the topical areas of information user related documents on Wikipedia. We found that Wikipedia covers various topics in the domain of information users. The LDA results uncovered "information retrieval" as a distinct area (Topic 2 in Table 2), and approximately 12.9% of the collected documents were assigned to that topic. Information retrieval has been one of the main topics across different disciplines such as the computer science and the library and information science. The high probability terms in Topic 2 well summarize the characteristics of information retrieval that involves user queries, information systems, and relevance judgement. In addition, Topic 11 and Topic 15 are subordinate areas of information retrieval. Topic 11 is about document indexing for text retrieval, and Topic 15 is related to information retrieval evaluation and relevance judgement, such as precision and recall. This implies that information retrieval is considered a significant area in the domain of information users. The LDA results also revealed that user experience design and usability (Topic 19) is a prevalent topic in the information users domain. We found that approximately 11.4% of documents were associated with the topics of user experience or usability. The terms concerning user experience design and usability test occurred together in Topic 9, which implies the close association between

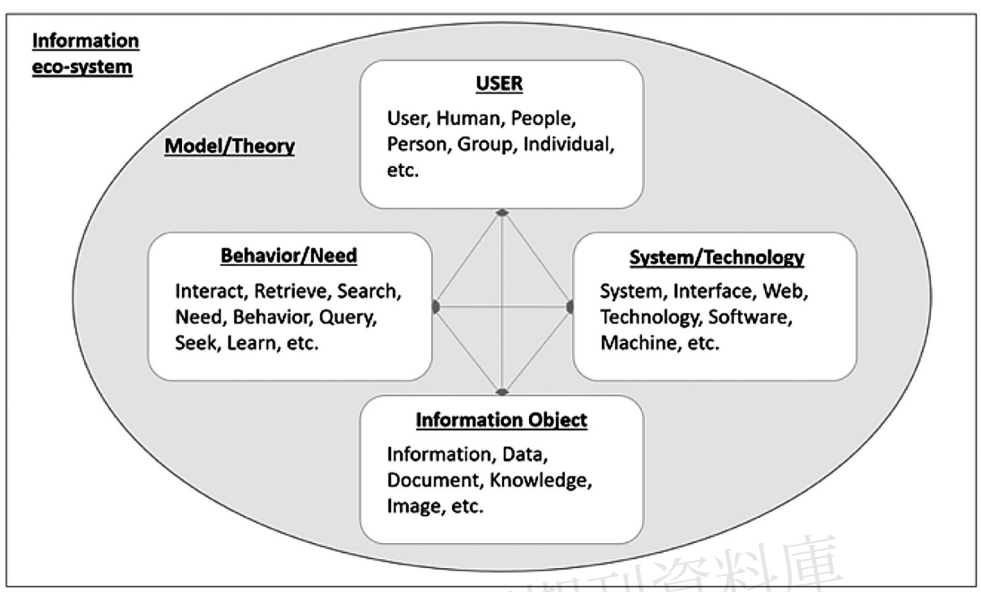


Figure 3. Key Components in the Domain of Information Users

臺灣大學學術期刊資料庫

the two concepts. Human-computer interaction was emerged as a separate topic (Topic 14). About 9.1% of the documents were relevant to human-computer interaction. We also could infer that human-computer interaction is closely related to interface design based on the interpretation of Topic 14. According to the topic models, web design and interface is also another important area of the information users domain (Topic 6). In addition, the LDA topic models detected multiple topics related to information seeking, information behavior, or associated theories and models. For example, Topic 1 includes the terms with high probability like "behavior" and "theori." Also, Topic 17 represents the topic of information seeking and behavior. The LDA topic models uncover several additional topics in the domain, such as social media, data mining and knowledge structure, and computer systems.

In addition, we looked into the relationships among the documents. Hierarchical clustering generated six large categories and twelve subordinate categories. Similar to the LDA analysis results, the largest clusters involve the topics of information retrieval and information search behavior. The first two clusters, C1 and C2, are related to information search behaviors. The C1 cluster includes a range of documents concerning users' search behavior, such as dwell time, faceted search, information foraging, browsing, and serendipity. User experience and design was identified as distinct clusters, i.e., C3 and C4. Particularly, C3 includes a set of documents directly relevant to user experience design, such as user experience, user interface design, heuristic evaluation, first time user experience, end user, and user story. C4 is closely related to C3, but C4 has more terms of "human" and "interaction." C5, C6, C7, and C8 represent the topics of information seeking and retrieval. Image retrieval related documents are clearly categorized into C5. C7 has a coherent set of documents concerning information retrieval. One noticeable finding is that the clustering analysis made a separate group of theory related documents (C12). In sum, we observed that the hierarchical clustering analysis results showed that a majority of the collected documents were clustered into information searching or information retrieval related categories. Also, user experience related documents formed another substantial cluster. Some minor clusters indicate the topics of indexing, research disciplines, information retrieval evaluation, human factors, and theories.

The dimensional reduction method, i.e., PCA, complemented the findings from the clustering analysis. It well portrayed the entire sphere of the information users domain. Using PCA, we defined 12 specific dimensions of the domain, ranging from information seeking and searching, user interface, user experience, information retrieval, human factors, and several others. The result of PCA was similar to that of the hierarchical clustering, but the PCA result made three substantial dimensions more distinctly, including human factors (D5), web and data mining (D10), and customer experience (D11). One of the important implications of the PCA results lies in the findings that many of the documents were associated with multiple topics. A fair number of documents were loaded to multiple components, indicating possible correlations across the topics. For example, the results revealed the close relationship between User interface (D2) and

臺灣大學學術期刊資料庫

User experience (D3). Also, some associations were detected between User experience (D3) and Human factors (D5) as well as between Human factors (D5) and Human computer interactions (D6). In sum, the PCA results suggest that documents were likely to be associated with each other, revealing the connected nature of Wikipedia articles under the higher theme of information users.

Finally, this study examined co-occurrences of terms based on the visualization method. The term co-occurrence map clearly illustrates that "user," "information," and "system" would be the three most important concepts in the domain of information users. These three key terms tend to have more connections with other terms. That is, these three concepts – users, information, and systems – lay the foundation for the domain of information users. We further investigated which concepts were relevant to each of these key terms: First, the term "user" was likely to have more links with the terms of design, user experience, method, experience, persona, usability testing, ergonomic, among others. Second, the term "system" has more chances to occur with the terms of computer, interface, time, computing, memory, implementation, and operating system. Third, the term "information" has higher probabilities of connection with the terms like information science, library, field, technology, and discipline. Other prominent terms emerged from the co-occurrence analysis include theory, behavior, research, document, use, and person. Particularly, the terms of research, use, theory, need, study, behavior, and factors were prone to occur together, which describe the areas of user behavior research, user needs, and associated theories. In this way, the term co-occurrence analysis disclosed the structure of linked terms that highlighted the three key concepts, including "user," "system," and "information," and then, presented closely related terms for each of the key concepts.

One of the contributions of this study lies in that it employed existent text mining methods widely applied in bibliometrics studies to explore topics and key concepts in the domain of information users by analyzing Wikipedia articles. We tried to discover a wide variety of topics and concepts by applying existing textual analysis methods. To do that, we thoroughly collected Wikipedia articles related to information users. Then, we investigated terms and topics of the domain, and also examined the relationships among the documents. The findings of this study yield insight into the holistic understanding of the domain of information users. Methodologically, this study employed existent techniques in text mining to better explore different aspects of the domain.

This study has several limitations. Even though we tried to include as many relevant Wikipedia documents as possible in the user research domain, the data set used in this study might not be comprehensive enough to depict the entire domain of information users. There are lots of recent research articles produced in many academic venues, but Wikipedia might not reflect all most up-to-date knowledge. In addition, the result of term frequency analysis tended to include more terms appeared in the seed topic documents amongst highly ranked terms. The selection of the seed documents might have a biased impact on the results. Moreover, this study did not suggest any novel method for Wikipedia mining. The study

臺灣大學學術期刊資料庫

adopted existing methods that are widely used in analysis of unstructured text data. To address this limitation, we plan a next study that involves a large set of most recent academic articles, such as journal articles and conference papers, in the domain of information user research. In addition, more elaboration will be sought to collect data that better represent a domain while minimizing any potential bias.

References

Alben, L. (1996). Quality of experience: Defining the criteria for effective interaction design. *Interactions, 3*(3), 11-15. doi: 10.1145/235008.235010

Azer, S. A. (2014). Evaluation of gastroenterology and hepatology articles on Wikipedia: Are they suitable as learning resources for medical students? *European Journal of Gastroenterology & Hepatology, 26*(2), 155-163. doi: 10.1097/MEG.0000000000000003

Bawden, D. (2006). Users, user studies and human information behaviour: A three-decade perspective on Tom Wilson's "On user studies and information needs." *Journal of Documentation, 62*(6), 671-679. doi: 10.1108/00220410610714903

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84. doi: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56*(1), 71-90. doi: 10.1108/EUM0000000007110

Borlund, P. (2003) The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3). Retrieved from http://informationr.net/ir/8-3/paper152.html

Callan, J., Allan, J., Clarke, C. L. A., Dumais, S., Evans, D. A., Sanderson, M., & Zhai, C.-X. (2007). Meeting of the MINDS: An information retrieval research agenda. *SIGIR Forum, 41*(2), 25-34. doi: 10.1145/1328964.1328967

Chen, B., Lin, Z., & Evans, T. S. (2019). *Analysis of the Wikipedia network of mathematicians*. Retrieved from https://arxiv.org/abs/1902.07622

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. *First Monday, 11*(11). doi: 10.5210/fm.v11i11.1413

Da Silva, T. S., Silveira, M. S., Maurer, F., & Hellmann, T. (2012). User experience design and agile development: From theory to practice. *Journal of Software Engineering and Applications, 5*(10), 743-751. doi: 10.4236/jsea.2012.510087

Everitt, B. S., Landau, S. & Leese, M. (2001). *Cluster analysis* (4th ed.). New York, NY: Oxford University Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438*, 900-901. doi: 10.1038/438900a

Halavais, A., & Lackaff, D. (2008). An analysis of topical coverage of Wikipedia. *Journal*

臺灣大學學術期刊資料庫

of Computer-Mediated Communication, 13(2), 429-440. doi: 10.1111/j.1083-6101.2008.00403.x

Harman, D. K., & Voorhees, E. M. (2006). TREC: An overview. *Annual Review of Information Science and Technology, 40*(1), 113-155. doi: 10.1002/aris.1440400111

Heilman, J. M., & West, A. G. (2015). Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language. *Journal of Medical Internet Research, 17*(3). doi: 10.2196/jmir.4069

Hollingsed, T., & Novick, D. G. (2007). Usability inspection methods after 15 years of research and practice. In D. Novick (Chair), *Proceedings of the 25th Annual ACM International Conference on Design of Communication* (pp. 249-255). New York, NY: Association for Computing Machinery. doi: 10.1145/1297144.1297200

Iba, T., Nemoto, K., Peters, B., & Gloor, P. A. (2010). Analyzing the creative editing behavior of Wikipedia editors: Through dynamic social network analysis. *Procedia - Social and Behavioral Sciences, 2*(4), 6441-6456. doi: 10.1016/j.sbspro.2010.04.054

International Organization for Standardization. (1998). *ISO 9241-11:1998*(en). *Ergonomic requirements for office work with visual display terminals (VDTs) —Part 11: Guidance on usability*. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en

Ingwersen, P., & Järvelin, K. (2006). *The turn: Integration of information seeking and retrieval in context*. Dordrecht, Netherlands: Springer.

Jackson, J. E. (2005). *A user's guide to principal components*. Hoboken, NJ: John Wiley & Sons.

Jatowt, A., Kawai, D., & Tanaka, K. (2016). Digital history meets Wikipedia: Analyzing historical persons in Wikipedia. In N. R. Adam, L. Cassel, & Y. Yesha (Chairs), *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 17-26). New York, NY: Association for Computing Machinery. doi: 10.1145/2910896.2910911

Jeng, J. (2005). Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *Libri, 55*(2/3), 96-121. doi: 10.1515/LIBR.2005.96

Joo, S., & Lee, J. Y. (2011). Measuring the usability of academic digital libraries: Instrument development and validation. *The Electronic Library, 29*(4), 523-537. doi: 10.1108/02640471111156777

Kipp, M. E. I., Joo, S., & Choi, I. (2013, August). *Classification of online resources using user generated terms*. Paper presented at IFLA WLIC 2013 - Singapore - Future Libraries: Infinite Possibilities in Session 112 - Classification and Indexing, Singapore.

Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, & G. M. Olson (Eds.), *Proceedings of the SIGCHI*

臺灣大學學術期刊資料庫

Conference on Human Factors in Computing Systems (pp. 1509-1512). New York, NY: Association for Computing Machinery. doi: 10.1145/1518701.1518930

Kräenbring, J., Penza, T. M., Gutmann, J., Muehlich, S., Zolk, O., Wojnowski, L., ... Sarikas, A. (2014). Accuracy and completeness of drug information in Wikipedia: A comparison with standard textbooks of pharmacology. *PLOS ONE, 9*(9), e106930. doi: 10.1371/journal.pone.0106930

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361-371. doi: 10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#

Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information science*. Norwood, NJ: Ablex.

Lee, J. Y., & Kam, M. (2018). Exploration on possibility of the disciplinary convergence of the user studies and the research in practice. *Journal of the Korean Society for Information Management, 35*(1), 129-155. doi: 10.3743/KOSIM.2018.35.1.129

Maimon, O. Z., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook*. Heidelberg, Germany: Springer.

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies, 67*(9), 716-754. doi: 10.1016/j.ijhcs.2009.05.004

Mehdi, M., Okoli, C., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2017). Excavating the mother lode of human-generated text: A systematic review of research that uses the Wikipedia corpus. *Information Processing & Management, 53*(2), 505-529. doi: 10.1016/j.ipm.2016.07.003

Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence, 194*, 222-239. doi: 10.1016/j.artint.2012.06.007

Morse, E. L. (2002). Evaluation methodologies for information management systems. *DLib Magazines, 8*(9). doi: 10.1045/september2002-morse

Mühlhauser, I., & Oser, F. (2008). Does WIKIPEDIA provide evidence based health care information? A content analysis. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 102*(7), 1-7. doi: 10.1016/j.zefq.2008.06.020

Ni, X., Sun, J. T., Hu, J., & Chen, Z. (2009). Mining multilingual topics from Wikipedia. In J. Quemada & G. León (Chairs), *Proceedings of the 18th International Conference on World Wide Web* (pp. 1155-1156). New York, NY: Association for Computing Machinery. doi: 10.1145/1526709.1526904

Nielsen, J. (1993). *Usability engineering*. Cambridge, MA: Academic Press.

Nielsen, J. (1994). Usability inspection methods. In C. Plaisant (Ed.), *Conference Companion on Human Factors in Computing Systems* (pp. 413-414). New York, NY:

臺灣大學學術期刊資料庫

Association for Computing Machinery. doi: 10.1145/259963.260531

Nielsen, L., Hansen, K. S., Stage, J., & Billestrup, J. (2015). A template for design personas: Analysis of 47 persona descriptions from Danish industries and organizations. *International Journal of Sociotechnology and Knowledge Development (IJSKD), 7*(1), 45-61. doi: 10.4018/ijskd.2015010104

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology, 59*(6), 938-955. doi: 10.1002/asi.20801

O'Brien, H. L., & Toms, E. G. (2013). Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Information Processing & Management, 49*(5), 1092-1107. doi: 10.1016/j.ipm.2012.08.005

Rohrer, C. (2014). *When to use which user-experience research methods*. Retrieved from https://www.nngroup.com/articles/which-ux-research-methods/

Skov, M. (2009). *The reinvented museum: Exploring information seeking behaviour in a digital museum context* (Doctoral dissertation, Research Programme Information Interaction and Information Architecture, Royal School of Library and Information Science, Copenhagen, Denmark). doi: 10.13140/RG.2.1.3466.4167

Van Eck, N. J., & Waltman, L. (2011). *Text mining and visualization using VOSviewer*. Retrieved from https://arxiv.org/abs/1109.2058

Van Eck, N. J., & Waltman, L. (2019). *VOSviewer manual*. Leiden, Netherlands: Univeristeit Leiden. Retrieved from https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.10.pdf

Vakkari, P. (2000a, April). Cognition and changes of search terms and tactics during task performance: A longitudinal case study. In J. Mariani & D. Harman (Chairs), *Content-Based Multimedia Information Access* (Vol. 1, pp. 894-907). Paris, France: Le Centre de Hautes Etudes Internationales D'informatique Documentaire.

Vakkari, P. (2000b). Relevance and contributing information types of searched documents in task performance. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, & P. Ingwersen (Chairs), *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2-9). New York, NY: Association for Computing Machinery. doi: 10.1145/345508.345512

Vakkari, P. (2001). A theory of the task-based information retrieval process: A summary and generalisation of a longitudinal study. *Journal of Documentation, 57*(1), 44-60. doi: 10.1108/EUM0000000007075

Vošner, H. B., Kokol, P., Bobek, S., Železnik, D., & Završnik, J. (2016). A bibliometric retrospective of the journal computers in

臺灣大學學術期刊資料庫

human behavior (1991-2015). *Computers in Human Behavior, 65*, 46-58. doi: 10.1016/j.chb.2016.08.026

Wang, Y. (2018) *Analysis of family-health-related topics on Wikipedia* (Doctoral dissertation, University of Wisconsin-Milwaukee). Retrieved from https://dc.uwm.edu/etd/1945/

Wang, Y., Joo, S., & Lu, K. (2014). Exploring topics in the field of data science by analyzing wikipedia documents: A preliminary result. *Proceedings of the American Society for Information Science and Technology, 51*(1), 1-4. doi: 10.1002/meet.2014.14505101116

Wikipedia. (2019). *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia.

Wilkinson, D. M., & Huberman, B. A. (2007). Cooperation and quality in Wikipedia. In A. Desilets (Chair), *Proceedings of the 2007 International Symposium on Wikis* (pp. 157-164). New York, NY: Association for Computing Machinery. doi: 10.1145/1296951.1296968

Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation, 37*(1), 3-15. doi: 10.1108/eb026702

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation, 55*(3), 249-270. doi: 10.1108/EUM0000000007145

Wilson, T. D. (2000). Human information behavior. *Informing Science, 3*(2), 49-55. doi: 10.28945/576

Xie, I. (2008). *Interactive information retrieval in digital environments*. Hershey, PA: IGI.

Xie, I., Joo, S., & Bennett-Kapusniak, R. (2017). User involvement and system support in applying search tactics. *Journal of the Association for Information Science and Technology, 68*(5), 1165-1185. doi: 10.1002/asi.23765

Xu, Q. (2015, March). *Examining user engagement attributes in visual information search*. Paper presented at iConference 2015 Create Collaborate Celebrate, Newport Beach, CA.

Zhao, Y., Liu, J., Tang, J., & Zhu, Q. (2013). Conceptualizing perceived affordances in social media interaction design. *Aslib Proceedings, 65*(3), 289-303. doi: 10.1108/00012531311330656

(Received: 2020/4/1; Accepted: 2020/5/15)

臺灣大學學術期刊資料庫

圖書資訊學刊　第18卷第1期　（2020.6）　頁1-23
doi: 10.6182/jlis.202006_18(1).001

資訊「使用者」領域之探索：以維基百科條目語意分析為例

Exploring the Domain of Information “Users”: Semantic Analysis of Wikipedia Articles

Soohyung Joo[1]

摘　要

本研究透過文本分析維基百科中的條目，以探索資訊使用者領域包含之概念與主題。本研究蒐集132則與資訊使用者有關之維基百科條目，運用多種文本分析方法，包括：字詞頻率分析（term frequency analysis）、潛在Dirichlet分配主題模型（Latent Dirichlet Allocation topic models）、階層式分群法（hierarchical clustering）、維度縮減（dimensional reduction）、詞語共現分析（term co-occurrence analysis）等，探索與資訊使用者有關之各種概念與主題。研究結果顯示，維基百科在資訊使用者領域中包含多元主題詞條，包括4種基本概念：使用者、行為、系統／科技、資訊物件。此外，本研究也分辨出此領域中的不同範圍，包含資訊搜尋行為、資訊檢索、人機互動、使用者體驗與人因等。

關鍵字：資訊使用者、領域分析、文本分析、維基百科、使用者研究

[1] 美國肯塔基大學資訊科學系
School of Information Science, University of Kentucky, USA
E-mail: soohyung.joo@uky.edu

註：本中文摘要由圖書資訊學刊編輯提供。
以APA格式引用本文：Joo, S. (2020). Exploring the domain of information “users”: Semantic analysis of Wikipedia articles. *Journal of Library and Information Studies, 18*(1), 1-23. doi: 10.6182/jlis.202006_18(1).001
以Chicago格式引用本文：Soohyung Joo. “Exploring the domain of information ‘users’: Semantic analysis of Wikipedia articles.” *Journal of Library and Information Studies* 18, no. 1 (2020): 1-23. doi: 10.6182/jlis.202006_18(1).001

臺灣大學學術期刊資料庫