Investigating interdisciplinary knowledge flow from the content perspective of citances

Jin Mao
School of Information Management
Wuhan University
Wuhan, Hubei, China
maojin@whu.edu.cn

Shiyun Wang
School of Information Management
Wuhan University
Wuhan, Hubei, China
wangsy2@whu.edu.cn

Xianli Shang†
Business School
Xinyang Agriculture and Forestry Univ
Xinyang, Henan, China
47655282@qq.com

ABSTRACT

Interdisciplinary research is playing an important role in modern science. In recent years, a lot of studies have measured interdisciplinary knowledge flow based on the frequency of citations. However, this approach does not consider the content of knowledge carried in the citations. In this study, we attempt to investigate the content of knowledge flow towards an interdisciplinary field by analyzing the citation sentences (i.e., citances) in the articles of the field. An emerging field, eHealth, is chosen in the case study. The associated knowledge phrases between citances and the references of the field are identified and categorized to analyze the content and categories of knowledge spread from the source disciplines to the field. The result shows that the ranks of disciplines by the frequency of associated phrases are consistent with the ranks by the frequency of in-text citations. Distribution of associated phrases over categories and disciplines is also analyzed. The associated phrases of *research subject* are the most, followed by *entity*. This study contributes to the understanding of content characteristics about interdisciplinary knowledge integration.

CCS CONCEPTS

• Theory of computation ~ Semantics and reasoning ~ Program semantics ~ Categorical semantics; • Information systems ~ Information systems applications ~ Data mining; • Information systems ~ Information retrieval ~ Retrieval tasks and goals ~ Information extraction

KEYWORDS

Interdisciplinary research, Content classification, eHealth, In-text reference, Knowledge integration

1 Introduction

Interdisciplinary research has become an important research paradigm and many recent significant breakthroughs in science are the fruits of interdisciplinary research. One fundamental feature of interdisciplinary research is the integration of knowledge from multiple disciplines out of the field [1]. Methods, theories, tools and concepts from different disciplines are often integrated to solve complex research problems of interdisciplinary research. To understand the characteristics of interdisciplinary knowledge integration, citation analysis has often been used to examine knowledge flow among disciplines[2]. Conventionally, the knowledge flow to a field is simply measured by the number of references cited by the papers in the field. Different importance, motivations and many other aspects of citations in a paper are ignored.

Recent studies have shifted to investigate interdisciplinary knowledge flow from a finer-granular perspective by looking into the content and contexts of citations. Citation contexts have become more easily obtained in recent years, which embed the syntactic (e.g., the location of section and rhetoric style) and semantic (e.g., the meaning of citation content) information of citations[3]. Citation contexts have been used to differentiate the functions[4-5], importance[6] and knowledge contributions[7] of different citations. The rich information of citation contexts enables the analysis on what knowledge is integrated into an interdisciplinary field.

In this study, we attempt to explore the content of knowledge integrated into an interdisciplinary field, *eHealth*, by analyzing the citances. The field of eHealth is an emerging field, referring to all aspects of the intersection of health care and the Internet[8]. A citance that provides the context of a citation is denoted as the sentence that contains in-text reference information. Our research questions are what knowledge is integrated from the source disciplines to eHealth, and what types are the knowledge. In this study, we design an approach to analyze the content and categories of the knowledge shared between citances and the references. This study contributes to understanding the content characteristics of interdisciplinary knowledge integration.

2 Methodology

2.1 Data Collection

Two high impact eHealth journals, *Journal of Medical Internet Research* and *JMIR mHealth and uHealth*, were selected as our data sources. All 3,416 articles with XML files published from 1999 to 2018 were collected. We only focused on the 3,221 articles with the types of Original papers, Reviews and Viewpoints. The metadata of references, including title, citation type, journal name, DOI, PubMed ID, and publish year, were

†Corresponding author.

parsed from the XML files. Sentences were extracted by using the punctuations (periods, question marks, etc.) as sentence boundaries, then citances with in-text references were identified. In total, 115,456 citances and 140,572 reference records were obtained.

To complete the abstracts of references, the reference records were fetched by searching PubMed for PubMed ID or Web of Science (WoS) for DOI. In total, the abstracts of 89,649 reference records were collected.

2.2 Source Discipline Identification

To explore the source of input knowledge, the references were then categorized into the 22 disciplines of Essential Science Indicators (ESI). We used the 2018 version of ESI journal list that covers 11,727 journals with full titles, abbreviated titles and their disciplines they belong to.

We designed a pipeline to determine the ESI disciplines of the references. First, 7,393 distinct journal titles were obtained from the 104,888 reference records with the citation type of 'journal' and with DOI/PubMed ID. We manually completed the full titles for the abbreviated journal titles that cannot be found in the ESI journal list but with more than 2 references. Next, we identified the disciplines of references by matching their journal titles with the journal titles in ESI. However, there were still 8,393 reference records without the ESI discipline information. Since the coverage of journals in ESI is not as broad as in WoS journal list, the WoS subject categories were then used to infer the ESI disciplines of the journal titles that were not matched directly. We designed a method to map the WoS subject categories into the ESI disciplines. We calculated the likelihood of a WoS subject category belonging to an ESI discipline through its journals whose ESI disciplines are known. The ESI discipline with the highest probability was then determined as the ESI discipline of the WoS category. If a journal has multiple WoS subject categories, we also chose the ESI discipline that has the highest probability with all the WoS categories.

Finally, approximately 94.09% of journal reference records (98,685) get the discipline information.

2.3 Extracting and Classifying Associated Knowledge Phrases

Citation contexts contain information about the cited articles relevant to the citing papers[9-10]. We contempt that the words occurred in both citation context and the corresponding cited paper can reflect the explicit knowledge association between the two to a certain extent. In this study, we used the title and abstract to represent a cited paper (i.e., a reference) due to the difficulty of obtaining full text. We extracted noun phrases that carry meaningful concepts from the citances as well as the titles and abstracts of the references by using the package of spaCy, an open-source python natural language processing toolkit. Noun phrases with a single character or some wildcards (e.g., "#", "*", "@", etc.) were removed. So were those starting or ending with a number. Stop words listed in the NLTK package were also eliminated. Acronyms were identified and expanded into their full forms by using the scispaCy package. We used both the acronyms and their full forms in the matching process, but only retained the raw forms of the noun phrases extracted from the citances. Thus, an associated knowledge phrase is defined as a noun phrase appearing in both a citance and its reference, which could be regarded as the knowledge transferred from the reference to the citing paper.

To analyze the types of the knowledge that flows to the eHealth field, we designed a classification framework of associated knowledge phrases based on the previous studies [11-13]. Two graduate students familiar with the field of eHealth were recruited to annotate the categories of the associated knowledge phrases by following the steps:

1. Initializing knowledge classification framework. One author constructed a preliminary classification schema after reviewing the literature. Then the author randomly selected 100 knowledge phrases for trial annotation, organized the annotation details, and wrote an annotation specification document that provides detailed definition to each category with a few exemplar concepts.
2. Pre-annotation. Pre-annotation training was carried out for the two coders. Subsequently, two coders independently annotated 500 identical knowledge phrases randomly selected for pre-annotation. After labeling, we calculated the kappa statistics to assess the agreement of the two coders. The kappa was equal to 0.65, which was not as good as expected. Thus, two coders discussed the ambiguous cases with a professional in the eHealth field. We find some phrases may not make sense if they appear alone, but they are meaningful in the given context, therefore, there were many phrases that categorized into the research subject category or others category by different coders. After the discussion, two coders reached a consensus.

TABLE 1. The classification framework of associated knowledge phrases.

Category	Description	Exemplar phrases
Research subject	Subject terms related to research problem	idepression, diabetes, health information
Theory	Theory related phrases	TAM, social cognitive theory, transtheoretical model
Research methodology	Methodology used in research	systematic review, analysis, meta analysis, randomize control trial
Technology	Technique, device and system that used in research	mobile phone, web, smartphone, app
Entity	Human-related research object	patient, woman, child, adolescent
Data	Phrases related to dataset, data source and data material	twitter, qualitative datum, clinical datum
Others	Other phrases that cannot be	study, use, result, outcome, number, Canada, project,

	included in the above categories
	USA

3. Formal annotation. The two coders annotated all 24,132 unique phrases. During the annotation process, two coders maintained communication with the professional in the eHealth field to reach an agreement.

Our final framework contains seven categories, including research subject, theory, research methodology, technology, entity, data and others, which are defined in detail in Table 1.

3 Results

3.1 Dataset Description

We obtained 3,221 papers from the eHealth field with the publication year between 1999 and 2018. Some characteristics of our dataset for analysis are given in Table 2. In total, 115,456 citances and 98,685 reference records (55,744 distinct articles) with discipline information were extracted from our corpus. The 98,685 reference records were cited a total of 134,516 times (i.e., in-text references) in all citances. Roughly 90% of the reference records have abstracts.

TABLE 2. Characteristics of our dataset for analysis

Characteristics	Statistics
Citing papers	3,221
Citances	115,456
Reference records	98,685
Reference records with abstract	89,649
Unique reference articles	55,744
In-text references	134,516
In-text references with abstract	123,206

3.2 Source Disciplines

To address our research question, we analyzed the distribution of references over disciplines. Table 3 shows the number of unique cited articles, CountOne citations, and in-text citations for the 22 disciplines. The CountOne citations were obtained by counting each reference only once in a citing paper, whereas the in-text citations count all the mentions of references in the paper[14]. The disciplines are ranked by the number of unique references. It's observed that the ranks of the disciplines by CountOne citations are the same as the ranks by in-text citations. In the following analysis, we choose the top 10 disciplines with most unique references, which cover 96.95% of all unique references.

3.3 Distribution of Associated Knowledge Phrases over Disciplines

In total, 215,138 associated knowledge phrases were extracted between the citances and the 123,206 in-text references with abstracts. Here, we only analyze 211,454 knowledge phrases associated with the top 10 disciplines (98.29% of all). Table 4 presents the frequency of associated knowledge phrases by discipline. It should be noted that only references with abstracts were used to extract associated knowledge phrases, therefore, the numbers of in-text citations in Table 4 are different from those in Table 3. Clinical Medicine contains the most associated knowledge phrases, followed by Social Sciences, General and Psychiatry/Psychology. The ranks of disciplines by the frequency of associated knowledge phrases are in harmony with the ranks by the frequency of in-text citations.

TABLE 3. Distribution of references over source disciplines

Rank	Discipline	Unique references	CountOne citations	In-text citations
1	Clinical Medicine	24802	47968	66673
2	Social Sciences, General	12812	22530	30196
3	Psychiatry / Psychology	9371	15915	21606
4	Neuroscience & Behavior	1914	2414	3152
5	Multidisciplinary	1259	2052	2754
6	Computer Science	1153	1660	2278
7	Immunology	839	1185	1464
8	Economics & Business	693	949	1222
9	Biology & Biochemistry	632	1041	1398
10	Pharmacology & Toxicology	567	710	963
11	Agricultural Sciences	546	839	1145
12	Engineering	303	357	441
13	Molecular Biology & Genetics	254	323	425
14	Mathematics	181	271	312
15	Environment / Ecology	181	216	249
16	Chemistry	80	91	94
17	Microbiology	51	53	44
18	Plant & Animal Science	46	47	38
19	Physics	27	30	36
20	Geosciences	26	26	15
21	Materials Science	5	6	9
22	Space Science	2	2	2

In addition, we calculated the knowledge density in the flow (i.e., the average number of phrases per citation) through dividing the frequency of phrases by the number of citations for each discipline. On average, every citation from the disciplines carried more than one associated knowledge phrase. The scores of knowledge density are slightly different between the 10

disciplines. Pharmacology & Toxicology exceeds other source disciplines, with the most phrases per citation, while Computer Science contains the fewest phrases per citation.

TABLE 4. The frequency of associated knowledge phrases

Disciplines	Knowledge phrases	In-text citations	Knowledge density
Clinical Medicine	113,424	61,385	1.848
Social Sciences, General	46,532	28,008	1.661
Psychiatry / Psychology	31,765	19,446	1.633
Neuroscience & Behavior	5,365	3,014	1.780
Multidisciplinary	4,470	2,561	1.745
Computer Science	2,750	1,979	1.390
Immunology	2,434	1,352	1.800
Biology & Biochemistry	1,905	1,301	1.464
Pharmacology & Toxicology	1,620	876	1.849
Economics & Business	1,189	855	1.391

3.4 Knowledge Category Distribution among Source Disciplines

According to the annotation result, the number of associated knowledge phrases is shown for each category in Figure 1. The phrases in the category of *research subject* are the most, accounting for 43.8%. It shows that authors usually cite references related to their research subject. One noticeable thing is that there are many phrases in *others*, which is the second most. Such phrases often involve specific authors' names, geolocations, specific projects, funding and some meaningless phrases. These phrases are not subdivided in our classification framework. In addition, the categories of *entity* and *technology* have more phrases than *research methodology*. This result may be due to the field of our corpus is medical-related, the research in which requires the use of many medical instruments, and the research entities it targets often varies in terms of research subjects (e.g., different diseases).

Figure 2 presents the number of associated knowledge phrases in different categories over the disciplines. The knowledge category distribution over different disciplines is significantly different (Pearson Chi Square test, p-value < 0.001). The top 3 disciplines, Clinical Medicine, Social Sciences, General, and Psychiatry/ Psychology, supply the most numbers of phrases in all categories. For each discipline, most of the associated knowledge phrases are *research subjects*.

In general, the distribution of associated knowledge phrases in each discipline over the categories are similar to the overall distribution in the entire dataset. However, a few exceptions are also observed. The proportion of theory phrases over all the phrases in Economics & Business are much higher than that in other disciplines. Computer Science has a higher proportion of technology phrases comparing with other disciplines. This could be explained by that Computer Science provides the study of eHealth with a lot of technique support, and many eHealth research problems are related to Computer Science.

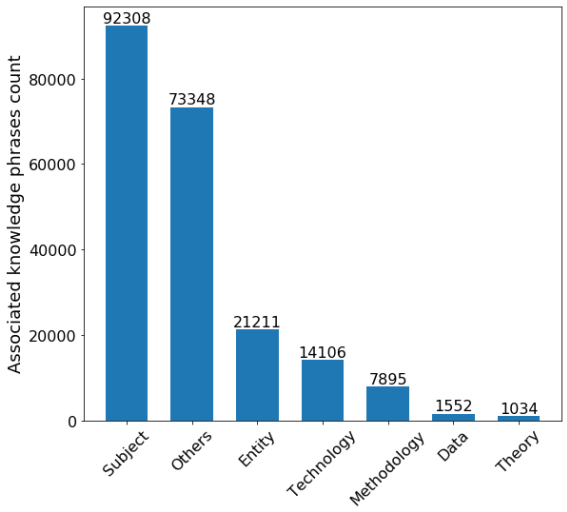

Figure 1: Frequency distribution of knowledge categories.

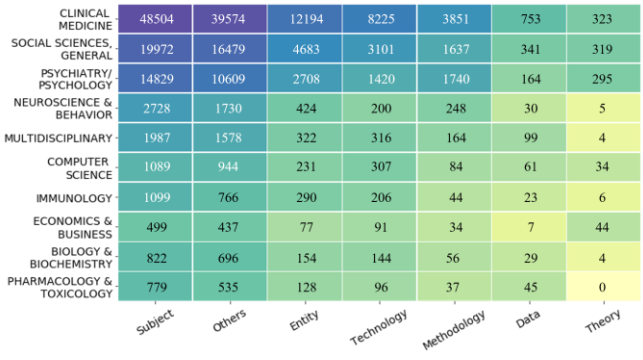

Figure 2: Frequency distribution of knowledge categories over disciplines.

4 Discussion & Conclusion

This study investigates the knowledge flow towards the interdisciplinary field of eHealth from the perspective of knowledge content. We extracted the knowledge phrases shared between the citances in the field with the references to represent knowledge content spread from source disciplines to the field. A classification framework was applied to annotate the identified knowledge phrases to explore the knowledge types of the phrases. The interdisciplinary features of eHealth are shown by analyzing the associated knowledge phrases.

The findings of this study could provide a few insightful implications on interdisciplinary knowledge integration. The result shows that the ranks of disciplines by the frequency of

associated phrases are consistent with the ranks by the frequency of in-text citations. It means that to measure interdisciplinary knowledge flow, an indicator based on the frequency of shared phrases may produce similar results with the indicator using the frequency of references, in that the in-text references from different disciplines often carry similar amounts of phrases (Table 4). Associated phrases can indicate the spread content, which may be useful to generate knowledge map of interdisciplinary knowledge integration. However, they do not directly differentiate citations, thus, it is not enough to only consider phrase frequencies to measure interdisciplinary knowledge integration at the aspect of content.

The frequency distribution of knowledge phrases over the categories is heavily skewed. Except *others*, the most in-text references carry the phrases of *research subject*, followed by *entity*. The results show the distribution of different types of knowledge from the source disciplines. The types of knowledge phrases can be used as an important feature to differentiate references, for instance, the motivations of citations. The categories of knowledge will be helpful to understand the roles of source disciplines in the knowledge integration of an interdisciplinary field.

A few limitations can be identified as well. To obtain full text of research articles, we only chose the two open access journals to represent the field of eHealth, which may not cover all the articles of this field. The problem of data deficiency is common in full-text based domain analysis. To identify the knowledge transferred from source disciplines to the interdisciplinary field, shared phrases are extracted by using simple text matching. However, synonyms are often used in citing others' work, thus the coverage of the shared knowledge may be in short.

We also identified some directions of future research. We manually annotated the categories of associated phrases. To support the analysis on large scale datasets, automating the classification of spread knowledge is on great demand, which is a challenging task of our interest. This motivates us to design a more general classification framework to analyze the content of knowledge spread between disciplines. In addition, recent machine learning techniques will be applied to this task in our future study.

ACKNOWLEDGMENTS

This study was funded by the National Natural Science Foundation of China (Grant No. 71804135) and Ministry of Education Humanities and Social Sciences project in China (Grant No.19YJC870018). We also thank Jing Tang for helping us with the data processing.

REFERENCES

[1] Porter A L, Cohen A S, Roessner J D, et al. 2007. Measuring researcher interdisciplinarity. Scientometrics 72(1),117-147.

[2] Yan E. 2016. Disciplinary knowledge production and diffusion in science. Journal of the Association for Information Science and Technology 67(9), 2223-2245.

[3] Ding Y, Zhang G, Chambers T, Song M, Wang X, and Zhai C. 2014. Content - based citation analysis: The next generation of citation analysis. Journal of the association for information science and technology 65(9), 1820-1833.

[4] Zhang G, Ding Y, and Milojević S. 2013. Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. Journal of the Association for Information Science and Technology 64(7), 1490-1503.

[5] Zhu X, Turney P, Lemire D, and Vellino A. 2015. Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology 66(2), 408–427.

[6] Hassan S U, Safder I, Akram A, and Kamiran F. 2018. A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. Scientometrics 116(2), 973-996.

[7] Thelwall M. 2019. Should citations be counted separately from each originating section?. Journal of Informetrics 13(2), 658-678.

[8] Pagliari C, Sloan D, Gregor P, Sullivan F, Detmer D, Kahan J P, ... and MacGillivray S. 2005. What is eHealth (4): a scoping exercise to map the field. Journal of medical Internet research 7(1), e9.

[9] Small H. 1978. Cited Documents as Concept Symbols. Social Studies of Science 8(3), 327-340.

[10] Elkiss A, Shen S, Fader A, Erkan G, States D, and Radev D. 2008. Blind men and elephants: What do citation summaries tell us about a research article?. Journal of the American Society for Information Science and Technology 59(1), 51-62.

[11] Wang, Y. and Zhang, C. 2018. What type of domain knowledge is cited by articles with high interdisciplinary degree? In *Proceedings of the Association for Information Science and Technology* 55, 1, 919–921.

[12] Gupta, S. and Manning, C. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, 1-9.

[13] Radoulov, R. 2008. Exploring automatic citation classification (master's thesis). Waterloo, Ontario, Canada: The University of Waterloo.

[14] Ding Y, Liu X, Guo C, and Cronin B. 2013. The distribution of references across texts: Some implications for citation analysis. Journal of Informetrics 7(3), 583-592.