Investigating interdisciplinary knowledge flow from the content perspective of citations

Wuhan University
Jin Mao, Shiyun Wang (Presenter) and Xianli Shang
August 1, 2020
Conventionally, the knowledge flow to a field is \textit{simply measured} by the number of references cited by the papers in the field.

Different \textit{importance, functions} and other aspects of citations in a paper are ignored.
Introduction

- **Citation contexts** embed the syntactic (e.g., the location of section) and semantic (e.g., the meaning of citation content) information of citations.
- In this study, we attempt to explore **what knowledge** is integrated into an interdisciplinary field, **eHealth**, by analyzing the **citances** (i.e., the sentence that contains in-text citations).

Adolescence and young adulthood are defined by developmental processes that mark increased susceptibility to risk-taking behaviors, including substance use [1-4]. In tobacco control, prevention efforts have shifted from individual and group-level interventions to population-based approaches, including policy and mass media efforts to reduce the appeal and accessibility of tobacco products to young people [5]. Concurrently, state-level cannabis policies in the United States have aimed to liberalize the accessibility of cannabis products, though there have been few state-level prevention campaigns. Using national surveillance data across states has been the standard approach to evaluate the effects of these policies on youth and young adult perceptions and behaviors [6,7]. These evaluations, which use cross-sectional data over time, may mask more nuanced trends in individual-level changes in harm perceptions and behavior and have largely failed to address spillover effects on other substance use. Novel surveillance methods that follow individuals over time and capture awareness of substance use prevention policy and communication efforts may provide a stronger basis for their evaluation.
Methodology

Step 1 Data Collection: Data collection and parsing

Step 2 Data Processing:
- Source Discipline Identification
- Associated Knowledge Phrases Extraction and Classification

Step 3 Data Analysis:
- Statistical analysis on knowledge phrases
- Distribution Analysis of Associated Knowledge Phrases
Data Collection

Two high impact journals, Journal of Medical Internet Research (JMIR) and JMIR mHealth and uHealth, in the eHealth fields, were selected as our data sources.
Source Discipline Identification

- 2018 version of Essential Science Indicators (ESI) journal list were used to identify the disciplines of our reference journals.

- 7,393 distinct journal titles
- Manually compensated 2,561 journal full titles
- Matching reference journal titles with the ESI journal list
- Still 8,393 reference records without the ESI discipline information

Web of Science (WoS) subject categories were used to infer the ESI disciplines of the not matched reference records

Probability calculation

Finally, approximately 94% of journal reference records (98,685) get the discipline information
Reference records without ESI discipline but with WoS category
We defined the noun phrases that appeared in both a citance and its reference as **associated knowledge phrases**.

**Titles and abstracts of references**

**Text processing (e.g., lemmatization)**

**nouns phrases extracted by spaCy**

**words filtering (stop words, wildcats, number words)**

**Scispacy were used to expand acronyms**

**Associated knowledge phrases**
Initializing knowledge classification framework.
• a author constructed a preliminary classification schema based on literature review
• randomly selected 100 knowledge phrases for trial annotation, and wrote an annotation specification document

Pre-annotation.
• two coders independently annotated 500 identical knowledge phrases
• coder discussed the ambiguous cases with a professional in eHealth to reach an agreement after the annotation process

Formal annotation.
• two coders annotated all 24,132 unique phrases, respectively
• maintained communication with the professional to reach a consensus during labeling
<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Exemplar phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Subject</td>
<td>subject terms related to research problems, e.g., drugs, diseases, research areas</td>
<td>e.g., <em>information</em>, <em>depression</em>, <em>diabetes</em>, <em>health information</em></td>
</tr>
<tr>
<td>Theory</td>
<td>theory related phrases, e.g., specific names of theories, frameworks, laws, etc.</td>
<td>e.g., <em>TAM</em>, <em>social cognitive theory</em>, <em>transtheoretical model</em></td>
</tr>
<tr>
<td>Research Methodology</td>
<td>methodology used in research, including research methods, scales, guidelines, evaluation indicators</td>
<td>e.g., <em>systematic review</em>, <em>analysis</em>, <em>meta analysis</em>, <em>questionnaire</em>, <em>randomize control trial</em></td>
</tr>
<tr>
<td>Technology</td>
<td>technique, device and system that used in research</td>
<td>e.g., <em>mobile phone</em>, <em>web</em>, <em>smartphone</em>, <em>app</em></td>
</tr>
<tr>
<td>Human Entity</td>
<td>people or organizations that are targeted by the experiment</td>
<td>e.g., <em>patient</em>, <em>woman</em>, <em>child</em>, <em>adolescent</em></td>
</tr>
<tr>
<td>Data</td>
<td>phrases related to dataset, data source and data material</td>
<td>e.g., <em>twitter</em>, <em>qualitative datum</em>, <em>clinical datum</em></td>
</tr>
<tr>
<td>Others</td>
<td>other phrases that cannot be included in the above categories, e.g., geolocations, funding, or some meaningless phrases</td>
<td>e.g., <em>study</em>, <em>use</em>, <em>result</em>, <em>outcome</em>, <em>number</em>, <em>canada</em>, <em>project</em>, <em>USA</em></td>
</tr>
</tbody>
</table>
Main Result 1

- The ranks of disciplines by the frequency of associated knowledge phrases are in harmony with the ranks by the frequency of in-text citations.
- The scores of knowledge density are slightly different between the 10 disciplines.

TABLE 4. The frequency of associated knowledge phrases

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

- The frequency distribution of knowledge phrases over the categories is **heavily skewed**

- Except **others**, the associated phrases of **research subject** are the most, followed by **entity** and **technology**
Main Result 3

• The knowledge category distribution over different disciplines is **significantly different** (Pearson Chi Square test, p-value < 0.001)

• the proportion of theory phrases in **Economics & Business** is much higher than that in other disciplines

---

Figure 2: Frequency distribution of knowledge categories over disciplines
Discussion & Conclusion

◇ Implications

1. Associated knowledge phrases can indicate the spread knowledge content, which may be useful to generate a knowledge map of interdisciplinary knowledge integration

2. Knowledge categories will be helpful to understand the roles of different disciplines in the knowledge integration of an interdisciplinary field
Thanks