A Probabilistic Model For Co-Occurrence Analysis in Bibliometrics

Xiaobei Zhou  
China Medical University

Miao Zhou  
China Medical University

Desheng Huang  
China Medical University

Lei Cui (lcui@cmu.edu.cn)  
China Medical University

Research Article

Keywords: probabilistic model, co-occurrence analysis, MeSH

Posted Date: June 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-634136/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. 
Read Full License

Version of Record: A version of this preprint was published at Journal of Biomedical Informatics on March 1st, 2022. See the published version at https://doi.org/10.1016/j.jbi.2022.104047.
A probabilistic model for co-occurrence analysis in bibliometrics

Xiaobei Zhou¹⁺, Miao Zhou¹⁺, Desheng Huang², and Lei Cui³⁺,*

¹Institute of Health Sciences, China Medical University, Shenyang, Liaoning, Peoples R China
²School of Public Health, China Medical University, Shenyang, Liaoning, Peoples R China
³School of Medical Informatics, China Medical University, Shenyang, Liaoning, Peoples R China

*To whom correspondence should be addressed. Email: lcui@cmu.edu.cn
⁺these authors contributed equally to this work

ABSTRACT

The co-occurrence analysis of Medical Subject Heading (MeSH) terms in the bibliographic database is popularly used in bibliometrics. Practically for making the result interpretable, it is necessary to apply a certain filter procedure of co-occurrence matrix for removing the low-frequency items that should not appear in the final result due to their low representativeness for co-occurrence analysis. Unfortunately, there is rare research referring to determine a critical threshold to remove noise of data for co-occurrence analysis. Here, we propose a probabilistic model for co-occurrence analysis that can calculate statistical significance (p-values) of co-occurred items. With help of this model, the dimensionality of co-occurrence network could be conveniently reduced according to selection of different levels of p-value thresholds. The conceptual model framework, simulation and practical applications are illustrated in the manuscript. Further details (including all reproducible codes) can be downloaded from the project website: https://github.com/Miao-zhou/Co-occurrence-analysis.git.

Introduction

Generally speaking, co-occurrence analysis is an old technique to analyze paired data within a collection unit. In bibliometrics co-occurrence analysis is used to study the potential relationship of two bibliographic items representing a certain research topic that appear in the same dissertation. All the bibliographic items of collection could form a co-occurrence matrix; and in a co-occurrence matrix the more co-occurrence between two items, the closer their relationship is. The methods for co-occurrence analysis in bibliometrics have become mature during last two decades¹,². Base on these methods Cui et al.³ developed a new approach focusing on analysis of Medical Subject Heading (MeSH) terms of PubMed database⁴. This is denoted as MeSH term co-occurrence analysis. A typical workflow of MeSH term co-occurrence analysis usually includes: (i) extracting MeSH term data in PubMed database; (ii) constructing a co-occurrence matrix with the intersection of two items indicating the co-occurrence by specific tools (e.g., BICOMB³); (iii) analyzing the co-occurrence matrix via a certain metric (e.g., biclustering⁵).

MeSH term co-occurrence analysis has been widely used by many researchers to study conceptual work in different research fields. Dehdarirad et al.⁶ researched the status and trends of gender difference in higher education and science. Hu et al.⁷ conducted co-word analysis/co-occurrence analysis to analyze the correlation coefficient of keywords about Chinese library and information science in the Chinese Journal Full-Text Database. Carlos et al.⁸ studied the development and stability of public health in Colombia and Mexico. Assefa and Rorissa⁹ described the basic structure of STEM education and analyzed the core areas and the relationship between them by using co-word analysis. Leydesdorff et al.¹⁰ developed a program to separate the Chinese texts, which made it possible to analyze Chinese words by using co-occurrence analysis. Sun et al.¹¹ retrieved literature of WoSCC database from 2015 to 2019 about acute pancreatitis. Referring to MeSH term co-occurrence analysis, there is always an unavoidable and unsolved problem: how to determine the critical thresholds of co-occurrence network or matrix. Sometimes, it is not practical to keep all the items when the sample size of document is huge. To many zero-counts in co-occurrence matrix due to low-frequency terms would cause the co-occurrence matrix sparse. In addition, the large number of low-frequency items included can increase the difficulty of interpreting the results. Unfortunately, there is rare literature related to this issue. Some researchers directly use Donohue formula¹² that is particularly designed for dividing high-frequency words and low-frequency words of an article based on Zipf’s law¹³. However, the item frequency distribution of an article may be very different from that of a collection of amount of dissertations in a certain research field. Whether Donohue formula can be applied to the co-occurrence analysis to determine the threshold, there is no systematic research yet. According to experience,
most of researchers choose about 30-40 high-frequency items, generally not more than 100, so that the co-occurrence matrix can form 4-5 controllable sub-categories, and each sub-category contains appropriated numbers of terms, which can more abundantly show the key situations in the research field\textsuperscript{14,15}.

In the entire MeSH item co-occurrence analysis process, manual selection of the threshold would make the results lose objectivity. Worse, if an inappropriate was chosen, incomplete or even wrong results would be obtained. This situation does not only mislead the readers of the study, but also makes the entire MeSH co-occurrence analysis criticized. To address the above issue, we develop an approach based on the Hypergeometric distribution specifically designed for MeSH term co-occurrence analysis and implement it using R program\textsuperscript{16}. With help of the tailored test (one-sided Hypergeometric test), statistically significant co-occurred items could be detected. We benchmark the method performance on the simulated data and make two real data applications.

Methods

The probabilistic model of MeSH term co-occurrence

Firstly, it is necessary to introduce some basic definitions to explain the model framework. The basic form of MeSH term co-occurrence data ($X_{m \times n}$) is a binary MeSH term by PMID\textsuperscript{17} matrix:

$$X_{m \times n} = \begin{pmatrix}
    \text{pmid}_1 & \text{pmid}_2 & \ldots & \text{pmid}_j & \ldots & \text{pmid}_n \\
    0 & 1 & \ldots & 0 & \ldots & 1 \\
    1 & 0 & \ldots & 0 & \ldots & 0 \\
    \vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\
    1 & 0 & \ldots & 0 & \ldots & 1 \\
    \vdots & \vdots & \ldots & \vdots & \ddots & \vdots \\
    1 & 0 & \ldots & 0 & \ldots & 0
\end{pmatrix}
$$

where $X_{m \times n}$ can be transferred into a MeSH term by term co-occurrence matrix $Y_{m \times m}$ by $X^T X$. Then $Y_{m \times m}$ is represented as:

$$Y_{m \times m} = \begin{pmatrix}
    \text{term}_1 & \text{term}_2 & \ldots & \text{term}_j & \ldots & \text{term}_n \\
    k_{11} & k_{12} & \ldots & k_{1j} & \ldots & k_{1m} \\
    k_{21} & k_{22} & \ldots & k_{2j} & \ldots & k_{2m} \\
    \vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\
    k_{n1} & k_{n2} & \ldots & k_{nj} & \ldots & k_{nm}
\end{pmatrix}
$$

For $Y_{m \times m}$, $k_{ij}$ represents the co-occurrence of between two MeSH terms ($\text{term}_i$ and $\text{term}_j$). It is assumed following a Hypergeometric distribution, denoted by $K_{ij} \sim \text{Hyper}(k, N_i, N_j)$, with:

$$P(K_{ij} = k) = \frac {\binom{N_i}{k} \times \binom{N - N_i}{N_j - k}} {\binom{N}{N_j}},$$

$$\max(0, N - k) \leq k \leq \min(N_i, N_j).$$

(1)

where $N_i$ = number of MeSH term$\text{term}_i$ occurred in all the publications, $N_j$ = number of MeSH term$\text{term}_j$ occurred in all the publications, and $N$ = total number of publications (or pmid). The term $\binom{N_i}{k}$ represents selecting $k$ from $N_i$ publications containing MeSH term$\text{term}_i$. The term $\binom{N - N_i}{N_j - k}$ means selecting $N_j - k$ from $N - N_i$ publications that have MeSH term$\text{term}_j$ but not term$\text{term}_i$. Multiplying these two terms together gives the total number of $k$ that both have MeSH term$\text{term}_i$ and term$\text{term}_j$. The denominator $\binom{N}{N_j}$ represents
total $N_j$ that could be obtained out of $N$ publication. Thus Equation 1 is giving the proportion of the $N_j$ samples that also have MeSH term$i$ under the condition that the two MeSH terms co-occur at $k$ times.

Here the one-sided Hypergeometric test (denoted as h-test) is specially designed for testing statistically significant pairs of co-occurrence items. The null hypothesis $H_0$ is that two MeSH terms ($term_i$ and $term_j$) are independent or negatively associated, and alternative test $H_1$ is that they are positively associated. If p-value $(K > k) < \alpha$, $H_0$ is rejected. Then two MeSH terms are significantly positively associated. In this case, they are denoted as significant co-occurred items (COIs).

**The simulation framework**

We built a simulation framework that aims to accurately reflect the reality of MeSH term co-occurrence data. The basic idea is that the event of a MeSH term occurred in a publication is assumed as a Bernoulli process. The number occurrence of MeSH terms in all the publications positively associated with others is assumed following a multivariate Bernoulli joint-distribution. The number occurrence of terms with no association is assumed following a binomial distribution. To reduce the complexity of simulation, we set up bivariate correlation for positively correlated MeSH terms. In order to evaluate the performance of our model, we created a simulation containing the following options:

1. **nt**: number of true class labels (COIs),
2. **nf**: number of false class labels,
3. **nc**: number of publications,
4. **rho**: correlation of COIs
5. **pf**: probability of false class labels,
6. **pt1**: probability of pairwise co-occurred MeSH term$i$,
7. **pt2**: probability of pairwise co-occurred MeSH term$j$.

Under an appropriate setting of parameters above, a simulated dataset (MeSH term by PMID matrix) can be generated. Based on our simulation framework, we generated 3 simulations according to different numbers of nc (200, 300 and 400). All other parameters are in the same and a reasonable setting: $nt = 50$, $nf = 100$, $rho \sim U(0.3, 0.9)$, $pt1 \sim U(0.01, 0.05)$, $pt2 \sim U(0.01, 0.05)$ and $pf \sim U(0.0001, 0.01)$. To evaluate the performance of our model, type I and II errors are calculated.

**Real data applications**

We re-analyzed one bibliometric study of pelvic organ prolapse\(^1^8\). In original paper, Zuo et al. made a MeSH term co-occurrence analysis from 3294 publications during 2007 – 2016. We compared the new result based on our model to the original one. **Note** that the dataset related to this paper is denoted as POP dataset.

We also generated a large dataset related to breast neoplasms extracted from PubMed database using our developed tool (**pubMR**: [https://https://github.com/xizhou/pubMR](https://https://github.com/xizhou/pubMR)). This dataset (denoted as Breast dataset) contains 108066 publications during 2010 – 2020 on breast neoplasms that can be inquired in PubMed database using search strategy ‘Breast Neoplasms[MAJR] AND (2010/01/01[PDAT] : 2020/01/01[PDAT])’. Among these publications, 3700 ESI high-level publications (highly cited or hot publications in the field) are considered as the truth.

**Customized metric**

We developed a customized correlation plot particularly for h-test. Additionally, the network plot in which the nodes represent the MeSH terms and the edges represent the connection between them is employed.

**Results**

**The probabilistic model could provide an unprecedented high-precision solution for exploring MeSH term co-occurrence**

We used POP dataset to illustrate the application of probabilistic model for MeSH term co-occurrence analysis. The result is a high dimensional p-value matrix ($3192 \times 3192$) (Supplementary Table S1). For interpreting this result, we employ our
The customized correlation plot (Figure 1) presents many detailed information, including all the COIs with a p-value less than 0.05 (large dot), their corresponding significant levels (with different colors) and the highest COIs with a p-value equal to 0 (marked with a green star). From this figure, we found 48 COIs with p-value equals to zero and additional 92 COIs (0 < p-value < 0.05) also taking our interest. To explore the relationship among these MeSH terms, we built their network structure (Figure 2). In this figure the nodes represent the MeSH terms; the edges represent the connection between them; the number or thickness of the edges represents their significant levels; and the links/edges between all the highest p-value of COIs are colored by blue.

Our result is compared with the original research (18). The original authors directly applied Donohue formula (denoted as d-method) as a threshold of high-frequency items. They collected 55 highly-frequent MeSH terms based on the threshold (> 35). The Venn Diagram plot comparing their collected highly-frequent MeSH terms and the COIs produced by our h-test is shown on Figure 3. From the figure, only partial terms (15) are overlapped by both methods, since two methods are based on different strategy: h-test provides the statistical inference; while the threshold approach is essentially taking a value that exceeds a certain threshold. Our results refers to the most statistically significant pairs of co-occurrence terms; while their results are the most highly frequent items of data.

Figure 1. The customized correlation plot of score matrix (1 – p-value). Large dot represent the COIs points with p-value less than 0.05. Different colors of dot represent different significant levels.
The h-test obtains a balanced statistical score in the experiment

To highlight how well the h-test could detect COIs, we focused on evaluating the performance of our model on the simulated data. The type I and II error rates of the model for different simulated datasets under the nominal alpha level ($\alpha = 0.05$) are shown in Table 1. We find that as the sample size ($NC$) increases, the type I error rate also increases, but it is still below the nominal alpha level. The type II error rate is very large in the small sample (e.g., $NC = 200$), and decreases sharply as sample size increases. A reasonable explanation is that in the case of a small sample size, the probability model lacks statistical capabilities and cannot detect true co-occurrences from random associations. The type II error rate is also affected by setting of correlation of COIs ($rho$). The larger the correlation is, the lower the type II error rate is. It is worth noting that if the sample size is small and the correlation of true pairs is low (e.g., $NC = 200$ and $rho = 0.1$), the model would perform poorly. In general, the function of the probability model performed well in multiple simulations with many parameter settings.

In addition, the performance of h-test on high-dimensional real data (Breast dataset) is also evaluated. Figure 5 shows the ROC curves of h-test and d-method applied on Breast dataset. h-test achieves a good tradeoff between power (true positive rate) and achieved type I error (false positive rate). And d-method has a fixed performance in case of a single threshold ($T = 464.15$). The extreme case that h-test strictly controls type I error under condition p-value = 0 is constructed and evaluated. Figure 4 (a) shows comparison between the reduced network applied by h-test under p-value = 0 (left) and the truth network (generated by
Figure 3. The Venn Diagram of obtaining the number of items on h-test and d-method. The number 15 means items are overlapped by both methods.

Figure 4. Network comparison between h-test under condition p-value=0 (a) and d-method in case of a single threshold $T = 464.15$ (b) on Breast dataset with the truth generated by ESI high-level publications.
Figure 5. The ROC curves of applying h-test and d-method on Breast datasets.

ESI high-level publications (right)). True positive, false positive and false negative of h-test are colored by gray, brown and green. Similarly, comparison between the reduced network applied by d-method and the truth network is represented Figure 4 (b). Overall, h-test under p-value = 0 controls type I errors well (less brown links) but loss some detection power (more green links), while d-method obtains a high power with a higher type I error.

<table>
<thead>
<tr>
<th>nc</th>
<th>rho=0.1</th>
<th>Type I error</th>
<th>Type II error</th>
<th>rho=0.2</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.0276161</td>
<td>0.66</td>
<td>0.0287077</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.0343036</td>
<td>0.38</td>
<td>0.0334128</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.0377917</td>
<td>0.18</td>
<td>0.0380176</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.0402886</td>
<td>0.01</td>
<td>0.0407905</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Type I and II error rates of the probabilistic model for multiple simulated datasets under different settings of nc and rho given the nominal alpha level ($\alpha = 0.05$).

Discussion

Regarding the application of the h-test, there is a practical issue in determining the minimum sample size (or publications) that require sufficient statistical power. Based on the result of simulation (Table 1), we consider that 200 of publications should be minimum limitation of application of h-test and 500 or more should be optimal. For the performance of h-test, there is a delicate tradeoff between type I and type II error. We still remains an open question as to how exactly to set an appropriate threshold for achieving a high power while maintaining a low false negative rate for a small sample size.

Conclusion

In this manuscript, we developed a new probabilistic model and h-test for MeSH term co-occurrence analysis and studied in detail its model frame, real case study and statistical power based on the simulation. Overall, the model seems to be adequate
for the task of fitting the MeSH term co-occurrence data well. We fully illustrated its model framework. Our real case study shows that h-test could successfully detect significant COIs and our simulation provides that h-test has enough statistical power.

Acknowledgements (not compulsory)

Thanks to Professor Cui Lei for his guidance on the research. We thank Zhou Miao for collecting and analyzing data, and thank Professor Huang Desheng for his model framework concept.

Author contributions statement

Zhou Xiaobei conceived the experiment(s) and completed experiment(s), Zhou Miao collected data and analysed the results. Cui Lei guided the research. Huang Desheng provided the model framework concept. All authors reviewed the manuscript.

References

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.csv
- SupplementaryTable1Caption.txt