

УДК 004.67

Характеристики сети соавторов медицинских наук

Д. В. Ланде, И. В. Горбов, И. В. Балагура

Институт проблем регистрации информации НАН Украины, Киев

Резюме

Целью данной работы является определение основных характеристик и закономерностей сетей соавторов в медицине и выделение в ней важных узлов. Исследована сеть соавторов в области «Медицина. Медицинские науки» в реферативной базе данных «Украиника наукова». Показано, что использование характеристик центральности узлов позволяют определить авторов научных работ, вносящих значительный вклад в отрасль, являющихся руководителями научных школ. Предложен и протестирован усовершенствованный коэффициент степени центральности, который добавляет вес авторам с равнозначными связями соавторства.

Ключевые слова: реферативная база данных «Украиника наукова», сложные сети, сети соавторов, центральность, медицинские науки.

Клин. информат. и Телемед.
2013. Т.9. Вып.10. с.141–144

Введение

Оценка научной деятельности является актуальной задачей, как в Украине, так и в мире. В последнее время в наукометрии наиболее частым объектом исследования в качестве показателя научного взаимодействия является соавторство в научных публикациях.

Научная коммуникация является движущей силой развития науки в современном мире. Обмен знаниями, опытом, идеями, ресурсами, информацией; политические, экономические и индивидуальные факторы; повышение качества и продуктивности результатов и просто удовольствие от работы побуждают ученых сотрудничать. Постоянное увеличение количества совместных исследований фиксировалось за рубежом неоднократно.

Под понятием «научная коммуникация» большинство зарубежных авторов понимают международное сотрудничество в рабочей группе для выполнения отдельных проектов. Результативность научного сотрудничества чаще всего проявляется в написании общих научных работ, патентов на изобретения, выступлениях на конференциях. Поэтому анализ совместных научных работ отображает наиболее объективные и видимые результаты сотрудничества. Коммуникабельность и влияние на развитие отрасли отдельных ученых чаще всего исследуют с помощью построения сетей соавторов и расчета параметров их узлов [1]. Под научным взаимодействием часто понимают так же образование научных школ. Общеизвестно, что развитие науки в Украине основывается на феномене «научных школ».

Целью данной работы является определение основных характеристик и закономерностей сетей соавторов в медицине и выделение в ней важных узлов.

Характеристики сетей соавторов

Сети соавторов являются одним из примеров использования концепции сетевой парадигмы, которая является эффективным инструментом исследования сложных систем. Для расчета характеристик сети в целом используются такие параметры, как количество узлов, количество ребер, среднее расстояние между узлами, диаметр сети – наибольшее геодезическое расстояние в сети и др. Определение клик (подгрупп или кластеров, в которых узлы связаны между собой сильнее, чем с членами других клик), выделение компонент (частей сети, которые связаны внутри и не связаны между собой), поиск перемычек (узлы, которые при их извлечении приводят к распаду сети) являются актуальными задачами исследования сложных сетей [2].

Степень разделения сети на группы оценивается коэффициентом кластеризации, который отражает отношение количества связей между соседями к полному возможному количеству таких связей. Общий коэффициент кластеризации графа вычисляется, как:

$$C = \frac{1}{N} \sum_{i=1}^N \frac{E_i}{k_i(k_i - 1)}, \quad (1)$$

где N – количество вершин, k_i – количество связей i -го узла, E_i – количество узлов, соседних с i -м узлом и связанных напрямую. Чем ближе значение коэффициента к 1, тем больше вероятность кластерной структуры [3].

Одной из важных характеристик графа является модулярность элементов и графа в целом. Модулярность узла – это величина, которая оценивает плотность связей в связанной компоненте

по сравнению со связями между компонентами. В общем виде модулярность определяется, как:

$$Q = \sum_{i=1}^N (e_{ij} - a_i), \quad (2)$$

где e_{ij} – элемент матрицы смежности графа, равный отношению количества ребер, которые объединяют два сообщества i и j , к общему количеству ребер в сети, $a_i = \frac{\sum_{j=1}^N e_{ij}}{N}$ – отношение количества ребер, объединяющих вершины в сообществе i , к общему количеству ребер. Высокая модулярность сети свидетельствует о сильной связи в сообществах – кластерах и слабой связи самой сети [4].

Методы исследования

В теории сложных сетей существует несколько типов коэффициентов, характеризующих относительную важность узлов в графе, таких как: степень центральности (degree centrality), центральность, как посредничество (betweenness centrality), центральность, как близость (closeness centrality) и центральность по собственному вектору (eigenvector centrality).

Степень центральности оценивает с каким количеством других участников связана конкретная личность, что для сетей соавторства так же можно понимать, как степень научного взаимодействия. В простейшем случае – это степень конкретной вершины:

$$C_D(i) = \sum_{j=1}^N m_{ij}, \quad (3)$$

где m_{ij} , если вершина i связана с вершиной j , и $m_{ij} = 0$, в противном случае [5].

Степень центральности характеризует авторов с точки зрения коммуникабельности и может использоваться для прогнозирования продуктивности автора. Недостатками данной характеристики для определения коммуникабельности ученых является отсутствие учета весов ребер, то есть совместных публикаций.

В работе [5] предложен коэффициент для подсчета степени центральности во взвешенном графе отдельных вершин:

$$C_D^{\alpha\alpha}(i) = k_i^{(1-\alpha)} s_i^{\alpha}, \quad (4)$$

где учитывается $k_i = \sum_{j=1}^N m_{ij}$ – сумма связей

с другими вершинами и $s_i = \sum_{j=1}^N \omega_{ij}$ – сумма весов соответственных связей, α – коэффициент, который подбирается для каждого отдельного случая.

Центральность, как близость, определяет, насколько близко вершина находится по отношению к остальным. Если автор центральный по данной характеристике, тогда он находится в центре исследований определенного направления и может изучить его все стороны с помощью своих соавторов или через них быстрее других получить необходимые для этого знакомства:

$$C_C(i) = \left[\sum_{j=1}^N d(i,j) \right]^{-1}, \quad i \neq j, \quad (5)$$

где $d(i,j)$ – наименьшее расстояние между вершинами i и j [6].

Центральность, как посредничество, определяет вершину, которая связывает между собой подграфы. В понимании научного сотрудничества посредничество дает возможность определить авторов, которые образуют связь между научными школами:

$$C_B(i) = \sum_{j < k} g_{jk}(i), \quad i \neq j, k, \quad (6)$$

где $g_{jk}(i)$ – число наиболее коротких путей в графе, которые проходят через i -ую вершину [6].

Метрика центральности по собственному вектору вычисляется с помощью учета весов соседних узлов, что дает возможность оценить узлы, которые связаны с важными соседями и вычисляется, как нормализованное значение собственного вектора матрицы смежности. Так же существуют метрики центральности такие, как эксцентриситет, авторитетность (Authority) и алгоритм PageRank. Эксцентриситет – расстояние от вершины к наиболее удаленной вершине, отражает возможность доступа к узлу от остальных узлов сети. Алгоритм PageRank и авторитетность были спроектированы для оценки связей между интернет-сайтами [6].

Для сетей соавторов важно выделять авторов, которые являются основателями научных школ и публикуют работы, вызывающие интерес как внутри области знаний, так и среди междотраслевых специалистов. При исследовании сетей соавторов в реферативной базе данных были выявлены авторы, публикующих чаще всего материалы исследований с одним и тем же соавтором и изредка сотрудничают с другими [7]. Таким образом важно различать авторов, которые опубликовали одинаковое количество работ ω , имеют одинаковое количество связей m с другими авторами, но одни

авторы одинаково продуктивны со всеми соавторами, а у других большая часть работ написаны в соавторстве с отдельными учеными. Данную особенность можно учесть с помощью использования коэффициента h -degree.

Отдельной категорией определения важности узлов являются коэффициенты на основе индекса Хирша. H -degree узла x во взвешенной сети равен k , если k – наибольшее натуральное число, что x имеет как минимум k связей с как минимум k весами [8]. Можно утверждать, что ученый, который сотрудничает с выдающимися коллегами, большинство из которых можно считать известными, имеет достаточно высокую коммуникативную способность. Данный принцип используется в l -index, lw -index, и в объединяющем h – подобные индексы важности узлов сети c -index. C -index узла x определяется, как c – наибольшее целое число, что как минимум с соседних узла имеют степень вершины и веса ребер, которые соединены с x не менее, чем c [9].

При определении важности узлов сети соавторов следует учитывать коммуникабельность, значимость соседних узлов, но не следует терять информацию об общей продуктивности авторов, так как количество соавторов не влияет напрямую на эффективность научной работы. Определение важных узлов в сети является актуальной задачей и требует детального изучения предмета исследования, т. к. существует множество коэффициентов, которые отражают разноплановые характеристики вершин, а целесообразность их использования определяется только соответствием целям экспериментов.

Для объединения характеристик оценки коммуникабельности и продуктивности ученых, скорректируем коэффициент степени центральности во взвешенном графе:

$$SC(i) = \frac{\sum_j m_{ij} \sum_j \omega_{ij} \beta_{ij}}{(n-1) \sum_{i,j} \omega_{ij}}, \quad (7)$$

т. е. модифицированный коэффициент центральности, содержит сумму связей и их весов и пронормирован на общее количество весов и связей в сети в соответствии с (2) и (3). Так же введен коэффициент β , который отражает отклонение весов от среднего для отдельного узла. Данный коэффициент уменьшает влияние отдельных связей соавторов с большими весами по отношению к остальным:

$$\beta_{ij} = e^{-\frac{(\omega_{ij} - \bar{\omega})^2}{(\sum_{i,j} \omega_{ij})^2}}, \quad (8)$$

$\beta \in (0;1]$ и при увеличении отклонения весов от среднего для отдельного узла стремится к 0, а при равенстве среднего значения весов для узла и веса равен 1. Таким образом снижено влияние авторов, у которых большинство совместных статей соответствует отдельным связям соавторства.

Результаты и обсуждение

В работе проанализировано содержимое реферативной базы данных «Україніка наукова», построена и проанализирована сеть соавторов по медицине. При построении сети соавторов в области «Медицина. Медицинские науки» использовался файл реферативной информации на момент ноября 2012 г., который содержит приблизительно 440 000 записей. «Медицина. Медицинские науки» является отдельной — 4-й серией украинского реферативного журнала, которая содержит рубрики: методика и техника научно-исследовательской работы в медицине, организация охраны здоровья, гигиена, общая патология, медицинская микробиология и паразитология, общая диагностика, общая терапия, фармакология, внутренние болезни, хирургия и др. [10]. Для увеличения точности результатов учитывался только первый инициал авторов, так же буквы «і», «ї», «є» в фамилиях и инициалах заменялись на «и», «е», соавторы с единичными связями в данной работе не рассмотрены. Вес связей между авторами каждой публикации рассчитывался как обратно пропорциональный к количеству ее авторов.

В результате программной обработки и программных средств визуализации сетей соавторства Gephi, получена сеть соавторов, которая содержит 40 199 авторов и 173 379 связей. Диаметр графа, т.е. максимальное расстояние между двумя соавторами через других авторов, равен 16. Средняя длина пути между двумя узлами 5. Среднее количество ученых, с которыми сотрудничает один автор, либо средняя степень узла приблизительно равна 9, а средняя взвешенная степень одного узла приблизительно равна 4, либо среднее количество работ в соавторстве одного автора — 4. Фрагмент сети соавторов представлен на рис. 1.

Сеть представляет собой слабо связанные между собой подграфы. Разделение сети на группы оценивалось с помощью коэффициента кластеризации

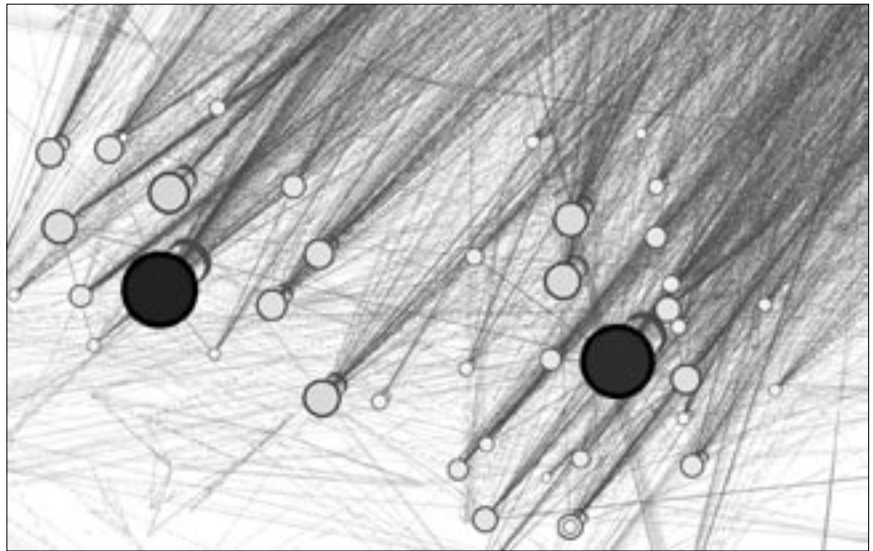


Рис. 1. Фрагмент сети соавторов в медицине и медицинских науках.

и модулярности. Модулярность сети соавторов равна 0,793, что свидетельствует об активном взаимодействии в мелких по сравнению с размером сети научных группах. Кластерная структура сети подтверждается так же коэффициентом кластеризации, который равен 0,771. В графе было найдено 1559 слабосвязанных компонент.

В работе были определены коэффициенты центральности узлов сети (табл. 1). В зависимости от результатов величина узлов сети была проранжирована. На рис. 1 изображена сеть соавторов с учетом расчета степени центральности. В результате расчетов и ранжирования узлов визуально проявились группы сильно связанных между собой авторов, которые сосредоточены вокруг одного или нескольких лидеров — больших узлов. Таким образом, визуально выявлены научные школы и коллективы организаций, которые тесно взаимодействуют в группе и через руководителей научных школ взаимодействуют с общей сетью. Усовершенствованный коэффициент степени центральности так же был протестирован и оказал влияние на первоначальное ранжирование авторов, добавив веса авторам с равнозначными связями соавторства. При ранжировании данных по усовершенствованному коэффициенту степени центральности, фамилии, объединяющие двух и более авторов однофамильцев с одинаковым инициалом, спустились ниже, по сравнению с предыдущими результатами, что дало возможность повысить ранг другим выдающимся ученым. Данные расчета коэффициента в табл. 1 помножены на 100, поскольку большой размер сети обуславливает высокую дробность результатов.

При определении различных коэффициентов важности узлов проявился недостаток наполнения реферативной базы данных в области медицинских наук. Несмотря на большой перечень медицинских журналов, реферируемых «Українікою науковою», часть публикаций остаются неосвоенными, т.е. недоступными для анализа. В результате некоторые выдающиеся ученые получили низкие оценки коэффициентов важности узлов. Методики на основе сетей соавторов являются достоверными и корректными лишь при полномочном представлении входных данных.

Выводы

В работе исследована сеть соавторов в области «Медицина. Медицинские науки». Определены основные параметры сети, такие как диаметр графа, общее и среднее количество узлов и связей, модулярность, коэффициент кластеризации и др.

Предложен и протестирован усовершенствованный коэффициент степени центральности, который добавляет вес авторам с равнозначными связями соавторства. Показано, что предложенный коэффициент понижает ранг фамилий, объединяющих двух и более авторов однофамильцев с одинаковым инициалом.

Показано, что полученная сеть соавторов представляет собой слабо связанные между собой подграфы, что обусловлено наличием большого количества различных медицинских наук и научных школ, недостатком представленности медицинских научных журналов.

Табл. 1. Ранжирование узлов сети соавторов по различным коэффициентом центральности (первые десять авторов).

Фамилия, имя	Центральн. как посреднич.	Фамилия, имя	Степень центральн.	Фамилия, имя	Усовершенств. коэффициент степ. центр.
Коваленко В.	2441523	Каладзе Н.	210,314	Бойко В.	0,18288
Бойко В.	2271441	Бойко В.	160,533	Чехун В.	0,14822
Москаленко В.	1949018	Ничитайло М.	151,907	Каладзе Н.	0,11791
Шевченко В.	1815715	Пишак В.	150,634	Пишак В.	0,11419
Чекман И.	1702083	Мишалов В.	149,725	Коваленко В.	0,10359
Дзяк Г.	1611266	Нейко Е.	136,767	Чекман И.	0,10264
Кравченко В.	1585579	Левицкий А.	121,421	Нейко Е.	0,09881
Грищенко В.	1472415	Чекман И.	121,28	Мишалов В.	0,09685
Пишак В.	1461707	Степаненко В.	112,991	Тронько М.	0,09653
Степаненко В.	1394735	Коркушко О.	103,807	Шевченко В.	0,08558

Большой размер сети соавторов по медицине и медицинским наукам обуславливает необходимость рассмотрения сетей соавторства отдельных более мелких отраслей. Так же существует необходимость включения полного перечня специализированных изданий по медицине и медицинским наукам, что увеличит достоверность исследований.

Литература

1. Bukvova, H. Studying Research Collaboration: A Literature Review [электронный ресурс] / Helena Bukvova // Sprouts: Working Papers on Information Systems. – 2010 – №10(3). – P. 1–19. – Режим доступа: <http://sprouts.aisnet.org/10-3>. – Назва з екрану.
2. Ландэ Д. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Снарский, И. В. Безсуднов – М.: Либроком (Editorial URSS), 2009. – 264 с.
3. Latapy M. Main-memory Triangle Computations for Very Large (Sparse (Power-Law)) / Matthieu Latapy // Graphs, in Theoretical Computer Science (TCS). – 2008 – V.407 №1–3, – P. 458–473.
4. Апанович З. В. Средства визуального анализа информационного наполнения порталов, входящих в облако Linked Open Data / З. В. Апанович, П. С. Винокуров, Т. А. Кислицина // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». RCDL/2011, Воронеж, Россия, 19–22 октября 2011 г. – CEUR, 2011 – V. 803 – P. 113–120.
5. Opsahl T. Node centrality in weighted networks: Generalizing degree and shortest paths / Tore Opsahl, Filip Agneessens, John Skvoretz // Social Networks. – 2010. – Vol.32, № 3. – P. 245–251.
6. Alireza A. Betweenness centrality as a driver of preferential attachment

in the evolution of research collaboration networks / Abbasi Alireza, Hossain Liaquat, Loet Leydesdorff // Journal of Informetrics. – 2012. – Vol. 6, № 3. – P. 403–412.

7. Ландэ Д. В. Наукометричні дослідження мереж співавторства по базі даних «Україніка наукова» / Д. В. Ландэ, І. В. Балагура // Реєстрація, зберігання і обробка даних. – 2012, – Т.14, №4 – с. 41–51.
8. Zhao S. X. H-Degree as a basic measure in weighted networks / Star X. Zhao, Ronald Rousseau, Fred Y. Ye // Journal of Informetrics. – 2011. – Vol.5, № 4. – P. 668–677.
9. Yan X. C-index: A weighted network node centrality measure for collaboration competence/ Xiangbin Yan, Li Zhai, Weiguo Fan // Journal of Informetrics. – 2013. – Vol.7, № 1. – P. 223–239.
10. Реферативна база даних «Україніка наукова» [електронний ресурс]. – Режим доступу: <http://nbuv.gov.ua/db/ref.html>. – Назва з екрану.

Characteristics in co-author network of medical sciences

D. V. Lande, I. V. Gorbov, I. V. Balagura
Institute for information recording
of NAS of Ukraine, Kiev

Abstract

The aim of the paper is to investigate main characteristics and features of co-author network of medical sciences and to recognize important nodes. Co-author network in areas of «Medicine. Medical sciences» of the Ukrainika naukova abstract database is studied in the work. There is shown that applying of nodes centrality characteristics allow to determine authors of scientific works which are heads of scholar schools, made a significant contribution and support communication between scientific teams. The improved degree centrality is proposed and tested. This coefficient modifies nodes weights with equal links.

Key words: the Ukrainika naukova abstract database, complex networks, co-author networks, centrality, medical sciences.

Характеристики мережі співавторів медичних наук

Д. В. Ландэ, І. В. Горбов, І. В. Балагура
Інститут проблем реєстрації інформації
НАН України, Київ

Резюме

Ціль даної роботи — визначити основні характеристики та закономірності мережі співавторів в медицині та виділити в ній важливі вузли. Було досліджено мережу співавторів в галузі «Медицина. Медичні науки» в реферативній базі даних «Україніка наукова». Показано, що використання характеристик центральності вузлів дозволяє визначити авторів наукових робіт, що вносять вагомий внесок у галузь, є керівниками наукових шкіл, підтримують зв'язок між науковими колективами. Запропоновано та протестовано вдосконалений коефіцієнт степені центральності, що додає ваги авторам з рівнозначними зв'язками співавторства.

Ключові слова: реферативна база даних «Україніка наукова», складні мережі, мережі співавторів, центральність, медичні науки.

Переписка

І. В. Балагура

Інститут проблем реєстрації
інформації НАН України
ул. Шпака, 2
Київ, 03113, Україна
тел.: +380 (44) 454 22 19
ел. пошта: balaguraira@mail.ru