УДК 612–08: 616

НОВЫЕ МЕТОДЫ АНАЛИЗА ПУБЛИКАЦИОННОЙ АКТИВНОСТИ НА ПРИМЕРЕ ИССЛЕДОВАНИЙ БОЛЕЗНИ ПАРКИНСОНА

Ф.Т. Алескеров
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: alesk@ipu.ru

А.В. Бульдяев
Московский физико-технический институт (Государственный Университет)
Россия, 141701, Московская обл., г. Долгопрудный, Институтский пер., 9.
E-mail: alexbuldyaev@gmail.com

О.Е. Хуторская
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: khutors@ipu.ru

А.И. Ямилов
Национальный Исследовательский Университет Высшая Школа Экономики
Россия, 101000, Москва Мясницкая ул., 20
E-mail: aibulatyamilov@gmail.com

Ключевые слова: сети цитирований, болезнь Паркинсона, базы данных публикаций

Аннотация: В работе любого исследователя постоянно возникает потребность в получении информации по тематике работы. Важно понимать, что уже сделано и что наиболее актуально в тематической области исследований. Такого рода информацию можно получить, используя различные базы знаний, объем которых увеличивается ежегодно. В связи с этим требуются новые методы их анализа, в частности, созданные в последнее время модели сетевого анализа. В работе предлагается комплексный подход к анализу публикационной активности, который использует методы сетевого анализа и тематического моделирования. Разработанный метод позволяет выявить связи между исследовательскими кластерами, ранжировать их значимость и отслеживать изменения направления исследований научных групп. Метод разрабатывался с применением к публикациям, посвященным различным аспектам болезни Паркинсона.

1. Введение

В последние годы наблюдается глубокий интерес к анализу различных сообществ и сложных сетей, особенно их структуры и ключевых элементов. Одним из применяемых методов анализа публикационной активности в мировой практике выступает анализ цитирований, то есть оценка влияния или качества исследования, автора, учреждения, журнала на основании количества цитирований другими исследователями. Публикаци-

онная активность и цитируемость может быть проанализирована методами сетевого анализа. Для этого публикации и их цитирования моделируются в виде графа, в котором публикации выступают в качестве вершин, а ребра графа несут информацию о цитированиях. Стоит отметить, что в ряде случаев графы достигают значительных размеров, как по количеству вершин, так и ребер. Для более детального изучения подобных сетей существуют методы, позволяющие выделить сообщества и изучить структуру сети. Альтернативным подходом к изучению сообщества является применение методов семантического анализа. Данные методы позволяют выделить тематическое направление каждой вершины сети, основываясь на неструктурированных текстовых данных.

В представленной работе используется комплексный подход с использованием сетевого, семантического и кластерного анализа в различных сообществах. В данной работе сообществами являются: публикации, журналы, авторы, аффиляции.

2. Материалы и методы

В качестве объекта исследования при создании новой методики сетевого анализа использовались публикационная и патентная активность в области исследования такого нейродегенеративного заболевания как болезнь Паркинсона (БП).

Болезнь Паркинсона и болезнь Альцгеймера (БА) – наиболее актуальные проблемы современной нейронауки. Рост числа больных, связанный с увеличением продолжительности жизни людей, привлекает к этой проблеме большое внимание исследователей, ученых и компаний в различных областях науки. Так, в базе данных научного цитирования Web of Science[1] хранятся библиографические записи о более 100 тысяч статей, опубликованных в 1980-2017 гг. и содержащих «parkinson» в ключевых словах.

При разработке метода использовалась база данных Web of Science (WoS). Рассматривались публикации по следующим идентификаторам тематики исследования: физиология, психология, психиатрия, хирургия, биохимия, фармацевтика и фармакология, химия, цитология, а также различные разделы биологии, биофизики, биотехнологии и биомедицины. WoS для анализа было отобрано более 75 тысяч публикаций за период с 1980 по 2017 гг., содержащих корень «parkinson» в (названии / аннотации/ ключевых словах). Публикации отбирались по 72 признакам, среди них: ID, наименование работы, аннотация, авторские ключевые слова, исходящие цитирования, авторы, ISSN, аффилиации, язык, год публикации и т.д.

2.1. Методика анализа

В качестве сети цитирований публикаций рассматривается граф, в котором вершины представляют работы (идентификационный номер публикации), а ребра представляют цитирования (связи) между ними. Такое представление позволяет проанализировать сети с помощью различных математических инструментов, выделить различные параметры графа, ключевые работы, а также дополнительные закономерности, представляющие возможный интерес для научного сообщества.

2.1.1. Индексы центральности. Для анализа использовались недавно разработанные алгоритмы оценки влиятельности элементов сети: индекс ближних взаимодействий Short-Range Interaction Centrality (SRIC[1]) и индекс дальних взаимодействий Long-Range Interaction Centrality (LRIC[2])). Ключевым преимуществом такого подхода по сравнению с существующими методами является то, что рассматривается как дальнее взаимодействие, так и специальные атрибуты вершин (в виде порогов), а также группо-

[1] https://clarivate.com/products/web-of-science/

вое влияние. Это позволяет обнаруживать скрытые ключевые вершины, которые влияют на прочие в группах или при дальних взаимодействиях.

Индекс ближних взаимодействий SRIC учитывает как прямое влияние одной вершины на другую, так и влияние через прочие публикации, имеющие прямую связь с данной. Алгоритм расчета состоит из 4 шагов:

1) Оценка интенсивности прямого влияния вершины i на вершину j;
2) Оценка интенсивности косвенного влияния вершины i на вершину j через вершину l, которая напрямую связана с i и j;
3) Оценка влияния вершины i на вершину j через все возможные вершины между ними с учетом группового влияния и порогового параметра;
4) Расчет финального значения индекса центральности для каждой вершины.

Индекс дальних взаимодействий LRIC является усовершенствованием индекса SRIC, он позволяет учитывать все возможные непрямые пути влияния одной вершины на другую. Оценка влиятельности вершин основана на анализе всех возможных цепей цитирований между ними. Алгоритм расчета индекса LRIC состоит из 5 шагов:

1) Оценка интенсивности прямого влияния вершины i на вершину j с учетом группового влияния и порогового параметра;
2) Для анализа непрямого влияния работы i на работу j рассматриваются все возможные простые цепи цитирований между ними (никакая работа в этой цепи не встречается дважды). С учетом того, что обычно очень длинные цепи цитирований не отражают реального влияния, вводится верхнее ограничение на длину пути;
3) Оценка интенсивности влияния вершины i на вершину j через каждую цепь из определенного перечня;
4) Определение финального влияния вершины i на вершину j путем агрегирования подсчитанных влияний для всех возможных каналов между вершинами;
5) Расчет финального значения индекса центральности для каждой вершины.

Таким образом, рассматриваемые индексы центральности позволяют выделить нетривиальные взаимосвязи между вершинами, определить скрытые влияния и выделить ключевые элементы с учетом вышеперечисленных свойств.

2.1.2. Семантический анализ (Тематическое моделирование). Большинство информации о публикациях: аннотации, ключевые слова представлены в текстовом формате. Увеличивающееся количество публикаций непрерывно наполняет объем доступных текстовых данных. Тематическое моделирование является основным инструментарием автоматического определения тем для больших коллекций документов. С помощью методов тематического моделирования можно анализировать коллекции научных статей и присваивать им темы, определять тенденции развития направлений исследований. В данной работе используется тематическая модель под названием Латентное размещение Дирихле (LDA) [3]. Данная модель предполагает, что каждый документ представляет собой совокупность тем, где тема – это набор слов, которые могут с разными вероятностями употребляться при обсуждении данной темы. На вход LDA подается коллекция документов, в данном случае коллекция аннотаций статей (или ключевых слов). На выходе модель для каждого документа определяет список тем с соответствующими вероятностями принадлежности документа той или иной теме. Таким образом, использование данного метода позволяет выделить ключевые тематики направления исходя из аннотации и авторских ключевых слов, представляющих из себя неструктурированные текстовые данные.

2.1.3. Выделение сообществ. В большинстве графов цитирований, а также в графах различных взаимодействий между агентами (в данном случае, исследователи или аффилиации) присутствует одна большая структурная компонента связности, в которой находится большинство вершин. Остальные компоненты содержат гораздо меньше

вершин. Структурный анализ такого графа является не информативным. Для более информативного анализа часто проводится выделение сообществ. Сообщества характеризуются тем, что вершины, входящие в одно сообщество, соединены между собой гораздо плотнее, чем с прочими вершинами графа. Выделение сообществ в графе способствует выявлению важной информации о графе и скрытых структурных особенностей. Существует ряд алгоритмов реализующих задачу определения сообществ. Почти все из них максимизируют модулярность – скалярную величину, которая обозначает разность между долей ребер внутри сообщества и ожидаемой долей связей при случайном размещении ребер. В данной работе для выделения сообществ был применен Multi-level modularity optimization algorithm – метод, основанный на многоуровневой оптимизации функции модулярности [4].

Алгоритм выделения сообществ в данном методе состоит из двух этапов:

Этап 1. В каждой вершине инициализируется сообщество. Для каждой вершины рассматривается возможность повышения модулярности за счет перемещения данной вершины в сообщество вершины-соседа. Выбирается наиболее выгодное перемещение с точки зрения значения модулярности. Данный шаг выполняется до тех пор, пока дальнейшее увеличение значения модулярности невозможно.

Этап 2. Создается новый граф с метавершинами в виде найденных сообществ и ребрами с суммарным весом всех ребер, идущих от одного сообщества к другому (так же появятся петли с суммарными весами связей внутри сообщества). Затем алгоритм перезапускается на образованном графе.

Таким образом, предлагаемый метод состоит из 3 пунктов:

a) Подготовка и обработка данных (информации о публикациях), включая семантический анализ аннотаций работ;
b) Построение и описание графов, включая применение алгоритмов выделения сообществ: граф цитирований публикаций, граф цитирований журналов, граф взаимодействия авторов граф цитирований аффилиаций, граф взаимодействия аффилиаций;
c) Анализ полученных сетей, включая расчет индексов центральности, ранжирование вершин графов (журналы или авторы, статьи или аффилиации) по влиятельности, поиск структурных взаимосвязей между статьями, темами исследований и аффилиациями, анализ развития направлений исследований.

3. Результаты

В качестве основного материала для анализа была использована следующая информация о статьях (в начале указано название соответствующего поля в нотациях WoS):

- TI: название статьи;
- AU: авторы статьи;
- SO: название журнала, в котором опубликована статья;
- SN: ISSN журнала;
- DE: ключевые слова, указанные авторами;
- ID: ключевые слова, присвоенные WoS;
- AB: аннотация статьи;
- C1: указание аффилиаций авторов статьи;
- PY: год публикации статьи;
- DI: уникальный присвоенный номер статьи DOI (Digital Object Identifier)
- WC: категория в рубрикаторе WoS;

- SC: область исследований.
- CR: указание цитирований на используемую литературу.

Формировались и рассматривались следующие типы графов:

a) Граф цитирований публикаций, позволяющий проанализировать взаимодействие и оказываемое влияние статей друг на друга;
b) Граф цитирований журналов - позволяет построить сеть взаимодействия научных изданий по определенной тематике;
c) Граф взаимодействия авторов, позволяющий рассматривать взаимодействие авторов в рамках публикации;
d) Граф взаимодействия аффилиаций - позволяет рассматривать взаимодействие научных центров, в которых работают авторы статьи и выделить центры, приносящие наибольший вклад в изучение тематики исследования.

Для корректного построения графа цитирований публикаций использовались DOI работ, так как данное поле присутствует в описании статей и идентификационные номера указаны в поле с цитированиями. Определение основных направлений исследований производилось с помощью применения алгоритма LDA к ключевым словам статей. Для более корректных результатов все слова были приведены к нижнему регистру, методика стемминга (выделение основы слова) не применялась в связи с риском потери смысла в специализированных терминах. На этапе обработки данных для каждой статьи были определены аффилиации исследователей с помощью обработки соответствующих полей в среде программирования Python. Проведенный анализ с использованием разработанного метода позволил выделить наиболее важные работы (авторов, аффилиации, журналы) в рамках тематики, ключевые и инновационные направления развития, а также провести ранжирование, отражающее значимость каждой работы. Также были выявлены исследовательские кластеры и связи между ними, и проведено их ранжирование по значимости.

4. Заключение

Основной целью данной работы являлось разработка и анализ возможностей применения системного подхода для изучения сетевых сообществ, а также выявления в них ключевых элементов. В работе рассмотрена возможность совмещения сетевого, семантического и кластерного анализа. Использование индексов ближних (SRIC) и дальних(LRIC) взаимодействий позволило в большей степени понимать причины влиятельности элементов. Разработанный метод позволяет исследователям получить информацию о ключевых направлениях исследований, развитии данных направлений, узнать о исследовательских кластерах и взаимосвязях между ними, выявить ключевых исследователей, аффилиации, журналы. Применение данного метода позволяет также отслеживать динамику популярности основных направлений исследований и привлекать внимание сообщества к возможно ранее неизвестным исследованиям.

Одним из преимуществ методики является гибкость применения различных методов сетевого и семантического анализа: результаты анализа можно варьировать, применяя другие алгоритмы выделения сообществ и расчета индексов центральности. Использование индексов ближних (SRIC) и дальних (LRIC) взаимодействий позволило в большей степени понимать причины влиятельности элементов. Одним из перспективных направлений развития работы является разработка конечного программного продукта, который позволит проводить анализ и отображать результаты по запросу пользователя в реальном времени.

Список литературы

1. Aleskerov F., Andrievskaya I., & Permjakova E. Key borrowers detected by the intensities of their short-range interactions // International Conference on Network Analysis. Springer, Cham. 2014. P. 267-280
2. Aleskerov F., Meshcheryakova N., Shvydun S. Centrality measures in networks based on nodes attributes, long-range interactions and group influence.2016. arXiv preprint arXiv:1610.05892.
3. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. No. 3. P. 993-1022.
4. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks // Journal of statistical mechanics: theory and experiment. 2008. No. 10. P10008. ArXiv ePrint: 0803.0476.
5. Алескеров Ф.Т., Бульдяев А.В., Хуторская О.Е., Ямилов А.И. Сетевой анализ публикационной активности и патентных баз по Болезни Паркинсона // Материалы 4-го Национального конгресса по болезни Паркинсона и расстройствам движений (Москва, 2017 г.). М.: ФГБУ «Научный центр неврологии» РАМН, 2017. Ч. 2. С. 295.
6. Buldyaev A., Aleskerov F., Khutorskaya O., Yamilov A. Parkinson's disease: network analysis of publications' activity // Polski Przegląd Neurologiczny. Warsaw: Via Medica Journals, 2018. Supplement A, Vol. 14. P. 146-147.