Факультет компьютерных наук

Прикладная математика и информатика

Москва 2023

Сетевые модели по публикациям в области научных исследований болезни Паркинсона

Network Analysis of Publications on Studies of Parkinson Disease

Работу выполнила:
Зиновьева Ксения, БПМИ194
Научный руководитель:
Алескеров Ф. Т.

Тематическая область

Болезнь Паркинсона (БП) - это нейродегенеративное, прогрессирующее заболевание, в основном характерное для людей старшего возраста.

В США в 2020 году 930 тысяч человек живут с БП, а к 2030 это число возрастет до 1.2 миллиона, каждый год болезнь диагностируется у 60 тысяч людей.

Прямые и косвенные затраты на БП в США составляют 52 миллиарда долларов ежегодно, на лекарства - $2500, а операции достигают стоимости в 100 тысяч долларов на одного человека.

Цели

- Проанализировать журналы и авторов в области исследований БП
- Применить новые модели анализа центральности
- Разработать целостную методику анализа научной области с использованием сетевых моделей

Обзор литературы

Анализ публикаций по БП:

- 100 самых цитируемых авторов сравниваются с помощью индекса Хирша [1],
- статистический анализ различных метрик для 100 самых цитируемых публикаций [2],
- анализ международных коллабораций, моделирование с помощью экспоненциальной регрессии для публикаций с 1991 по 2006 [3].

Сетевой анализ:

- Индексы центральности SRIC и LRIC в работе 2018 года [4],
- Сетевой анализ Российских экономических журналов [5].

Используя сетевой анализ цитирований, можно выделить ключевые исследования и журналы, в которых они публикуются.

Этапы работы

1. Обзор литературы и источников
2. Сбор данных из базы Microsoft Academic
3. Обработка данных
4. Предварительный анализ
5. Построение сети цитирования для журналов и авторов
6. Расчет индексов центральности по сетям
7. Выделение топ-10 вершин по индексам, их сопоставление
8. Анализ динамических изменений в индексах по годам
9. Основные выводы и направление дальнейших исследований
10. Подготовка статьи по данной теме

Описание данных

Из базы научных публикаций Microsoft Academic были скачаны статьи 2015-2021 года со словами “parkinson” и “disease” в названии или аннотации.

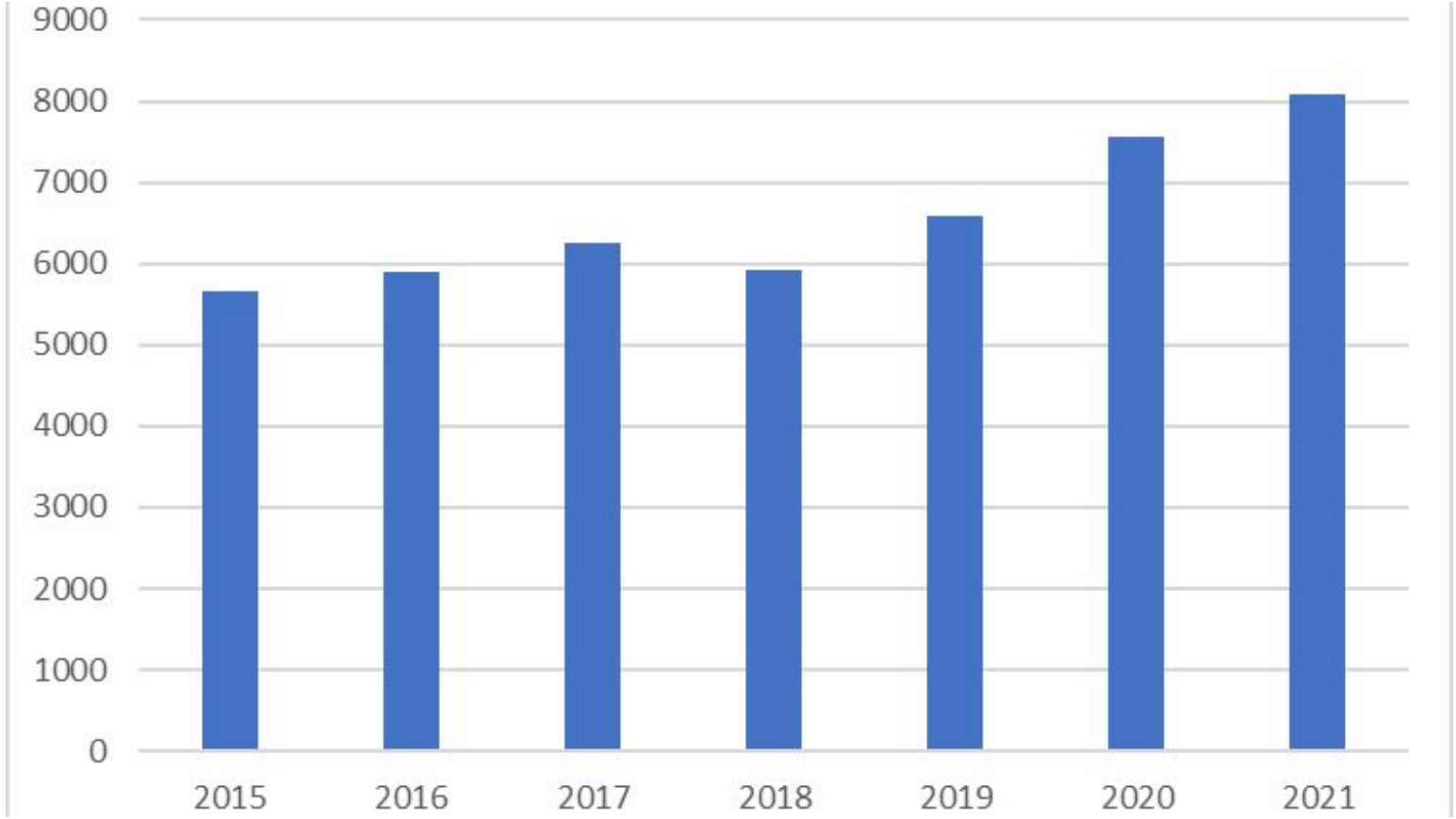
Атрибуты статьи:

- Id – ID публикации
- W, AW – уникальные нормализованные слова (приведенные к единой форме) в заголовке и абстракте
- Y – год публикации
- RId – список ID статей, на которые ссылается публикация
- DOI – цифровой идентификатор объекта
- J.JId – ID журнала

Всего статей скачано: 70119.
Из них

- **10681 без DOI**
- **7315 с DOI и без аннотации**
- **52123 с DOI и аннотацией**
- **45940 с DOI, аннотацией и журналом**

Распределение статей с DOI, аннотацией и журналом по годам



Индексы центральности

Классические индексы

1. In-degree index – сумма весов входящих ребер

$$x_i = \sum_j w_{ij}$$

2. Eigenvector index – решение уравнения $Ax = \lambda_1 x$, где λ_1 максимальное собственное значение матрицы смежности A

$$x_i = \frac{1}{\lambda_1} \sum_j A_{ij} x_j$$

Классические индексы

3. PageRank centrality – это разновидность центральности по собственному вектору, которая учитывает исходящую степень

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

4. Betweenness centrality – показывает долю кратчайших путей между двумя вершинами, на которых лежит исследуемая вершина

$$x_i = \sum_{kj} \frac{n_{kj}^i}{g_{kj}}$$

Новые индексы

2. Pivotal index (PI) показывает влияние ключевых вершин.

Вершина j_p называется ключевой, если $\sum_{j \in S} w_{ji} \geq q_i$ и $\sum_{j \in S \setminus \{j_p\}} w_{ji} < q_i$.

Значение индекса для критического множества равно количеству ключевых вершин в нем. Для вершины: $PI(i) = \sum_{S} |S| \times PI_i(S)$

Общее влияние: $TI(i) = \frac{1}{3} In - degree(i) + \frac{1}{3} BI(i) + \frac{1}{3} PI(i)$

Новые индексы

S – критическое множество для вершины i, если $S \subseteq V\backslash\{i\}$,

$|S| \leq k, \sum_{j \in S} w_{ji} \geq q_i$, где квота q_i - процент суммы весов входящих ребер, k – количество вершин, которые одновременно могут влиять на узел. В работе k=3. С увеличением квоты уменьшается количество критических множеств.

1. Bundle index (BI) учитывает групповое влияние на вершину

$$BI_i(S) = 1, \text{если} \sum_{j \in S} w_{ji} \geq q_i \qquad BI_i(S) = 0, \text{иначе}$$

$$BI(i) = \sum_{S} BI_i(S)$$

Aleskerov F., Yakuba V. Matrix-vector approach to construct generalized centrality indices in networks. 2020 // SSRN 3597948. Available at: https://ssrn.com/abstract=3597948

Сеть цитирований по журналам

Атрибуты сети:

- Jld1 – ID цитирующего журнала
- Jld2 – ID цитируемого журнала
- Y – год публикации журнала 1
- Weight – количество цитирований журнала 2 журналом 1 в год Y

Количество вершин в сети - 3292, количество ребер в сети 152203.

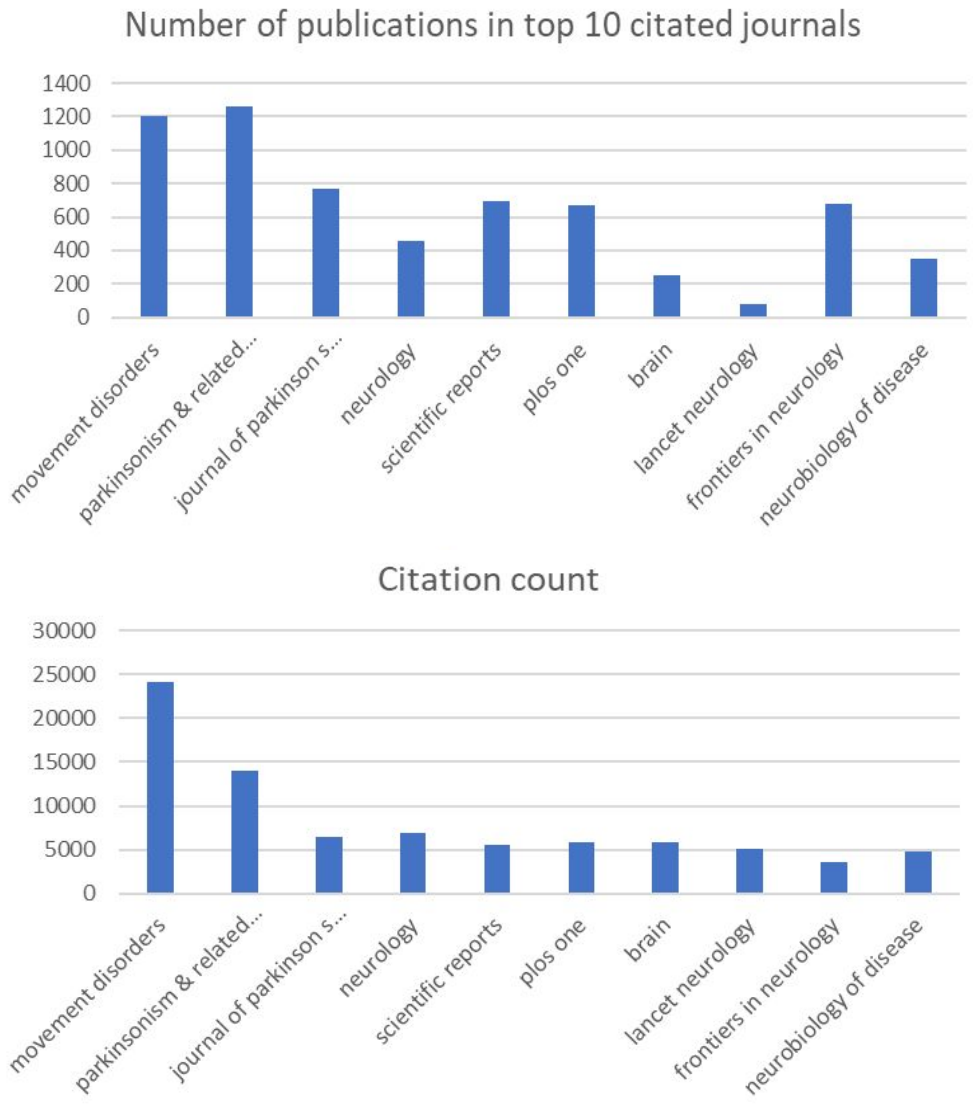
Jld1	Jld2	weight	Y
10623703	10623703	101	2021
10623703	163027424	150	2021
115201632	163027424	143	2020
115201632	163027424	134	2021
118428158	163027424	146	2020
118428158	163027424	127	2021
147691530	147691530	120	2017
147691530	147691530	247	2019
147691530	147691530	139	2020
147691530	147691530	195	2021
147691530	163027424	123	2016
147691530	163027424	115	2017
147691530	163027424	102	2018

Сеть цитирований по журналам

Количество вершин	3292
Количество ребер	152203
Количество компонент связности	4
Размер наибольшей компоненты связности	3285
Плотность графа	0.0094
Минимальное количество цитирований	1
Максимальное количество цитирований	24122
Среднее количество цитирований	113.5

Результаты для журналов

Name of journal	Number of citations
1. Movement Disorders	24122
2. Parkinsonism & Related Disorders	14035
3. Journal of Parkinson’s Disease	6524
4. Neurology	6851
5. Scientific Reports	5613
6. PLOS One	5911
7. Brain	5909
8. Lancet Neurology	5085
9. Frontiers in Neurology	3641
10. Neurobiology of Disease	4757



Самые цитируемые журналы

Топ-10 журналов по индексу TI, q = 0,5%

Название	In-degree	BI, q=0,5%	PI, q=0,5%	TI, q=0,5%
1. Movement Disorders	0,0472	0,0276	0,0247	0,0332
2. Parkinsonism & Related Disorders	0,029	0,0248	0,0248	0,0262
3. Journal of Parkinson’s Disease	0,0165	0,0168	0,0156	0,0163
4. PLOS one	0,014	0,0167	0,016	0,0156
5. Lancet Neurology	0,0137	0,0144	0,0145	0,0142
6. Scientific Reports	0,014	0,0129	0,0131	0,0134
7. Neurology	0,015	0,0105	0,0101	0,0119
8. Frontiers in Neuroscience	0,009	0,0114	0,014	0,0115
9. Frontiers in Neurology	0,011	0,0117	0,0113	0,01145
10. npj Parkinson’s Disease	0,0087	0,011	0,0127	0,0108

Топ-10 журналов по индексу TI, q = 5%

Название	In-degree	BI, q=5%	PI, q=5%	TI, q=5%
1. Movement Disorders	0,0472	0	0	0,0157
2. Parkinsonism & Related Disorders	0,029	0	0	0,0097
3. Age and Ageing	0,00033	0,008	0,0134	0,0072
4. British Medical Bulletin	0,0003	0,0067	0,0117	0,0062
5.Inflammopharmacology	0,00033	0,0071	0,011	0,0061
6. Expert Opinion on Drug Metabolism & Toxicology	0,0003	0,0063	0,0105	0,0057
7. Journal of Parkinson's Disease	0,0165	0	0	0,0055
8. Cells	0,0027	0,0087	0,005	0,00547
9. Genes & Development	0,0003	0,0059	0,0095	0,0052
10. Neurology	0,015	0	0	0,005

Топ-10 журналов по индексу TI, q = 10%

Название	In-degree	BI, q=10%	PI, q=10%	TI, q=10%
1. Movement Disorders	0,0472	0	0	0,0157
2. Parkinsonism & Related Disorders	0,029	0	0	0,0097
3. Environmental Toxicology and Pharmacology	0,00017	0,0062	0,0103	0,0056
4. Springer plus	0,00017	0,0062	0,0103	0,0056
5. Journal of Parkinson's Disease	0,0165	0	0	0,0055
6. Neuroimmunomodulation	0,00018	0,00626	0,0096	0,0053
7. Gastroenterology Research and Practice	0,00018	0,00626	0,0096	0,0053
8. Neurology	0,015	0	0	0,005
9. Amino Acids	0,00017	0,0056	0,0086	0,0048
10. Scientific Reports	0,014	0	0	0,0047

Сеть цитирований по авторам

Количество вершин	27551
Количество ребер	271623
Количество компонент связности	82
Размер наибольшей компоненты связности	27402
Плотность графа	0.0004
Минимальное количество цитирований	0
Максимальное количество цитирований	2416
Среднее количество цитирований	14.75

Топ-5 авторов по In-degree

Имя	Аффилиация	In-degree	Betweenness rank	Eigenvector rank	Pagerank rank
1. Ronald B. Postuma	Montreal General Hospital	0,0071	2	1	1
2.Anthony H.V. Schapira	University College London	0,0024	73	17	5
3.E. Ray Dorsey	University of Rochester	0,0023	19	62	3
4.Mike A Nalls	National Institutes of Health	0,002	207	10	10
5.Alberto J. Espay	University of Cincinnati	0,00185	6	12	24

Топ-5 авторов по TI q = 0.1%

Имя	Аффилиация	In-degree	BI, q = 0.1%	PI, q = 0.1%	TI, q = 0.1%
1.Ronald B. Postuma	Montreal General Hospital	0.0071	0,577	0,9999	0,528
2.E. Ray Dorsey	University of Rochester	0,0023	0,0251	0	0,0091
3.Anthony H.V. Schapira	University College London	0,0024	0,0245	0	0,009
4.Alicia M. Pickrell	National Institute of Neurological Disorders and Stroke (NINDS)	0,0018	0,0142	0	0,0053
5.Ole-Bjørn Tysnes	Haukeland University Hospital	0,00164	0,0122	0	0,0046

Топ-5 авторов по TI q = 10%

Имя	Аффилиация	In-degree	BI, q = 0.1%	PI, q = 0.1%	TI, q = 0.1%
1.Ronald B. Postuma	Montreal General Hospital	0,0071	0	0	0,0024
2.Thomas J. Hirschauer	Ohio State University	0,0001	0,0009	0,0015	0,0009
3.Yimeng Chen	Chinese Academy of Sciences	0,0001	0,0009	0,0015	0,0009
4.Tatsuya Sasaki	Okayama University	0,0001	0,0009	0,0015	0,0009
5.Michael Khalil	Medical University of Graz	0,0001	0,0009	0,0015	0,0009

Заключение

- Проанализировано почти 40 тысяч публикаций из более чем 3 тысяч журналов, почти 30 тысяч авторов
- Подсчитаны классические индексы и недавно разработанные Bundle index и Pivotal index
- Работа представлена на конференции "The 12th International Conference on Network Analysis" в мае 2022 года
- Работа представлена на конференции "HCist 2022 - International Conference on Health and Social Care Information Systems and Technologies" в ноябре 2022 года

Разработанная методика может применяться для выявления ключевых областей и участников исследований, в том числе для выгодных инвестиций.

Планы

1. Провести паттерн-анализ изменения индексов по годам для журналов
2. Исследовать устойчивость сетей цитирования журналов и авторов
3. Провести анализ аннотаций (семантическая близость публикаций)
4. Разработать модели по соавторам

Источники

[1] Sorensen, A. A., & Weedon, D. (2011). Productivity and impact of the top 100 cited Parkinson's disease investigators since 1985. Journal of Parkinson's disease, 1(1), 3–13.

[2] Xue, J. H., Hu, Z. P., Lai, P., Cai, D. Q., & Wen, E. S. (2018). The 100 most-cited articles in Parkinson's disease. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 39(9), 1537–1545.

[3] Li, T., Ho, Y. S., & Li, C. Y. (2008). Bibliometric analysis on global Parkinson's disease research trends during 1991-2006. Neuroscience letters, 441(3), 248–252.

[4] F. Aleskerov, O. Khutorskaya, A. Buldyaev, and A. Yamilov, "Parkinson's disease: Network analysis of publications' impact," IEEE Xplore, May 01, 2018.

[5] F. Aleskerov, D. Karabekyan, A. Kazachinskaya, A. Semina, and V. Yakuba. (2021) "Economic journals ofRussia, their characteristics and network analysis." Journal of the New Economic Association 50 (2): 170–182.

[6] F. Aleskerov and V. Yakuba, "Matrix-Vector Approach to Construct Generalized Centrality Indices in Networks," papers.ssrn.com, May 11, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract id=3597948

[7] Newman MEJ. (2010). Networks, an Introduction. New York, NY: Oxford University Press.

[8] Bonacich,P.(1972).Factoring and weighting approaches to status scores and clique identification. Journal of mathematical sociology, 2(1), 113-120.

[9] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117.

[10] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry, 35-41.

Спасибо за внимание!