ПАРАМЕТРЫ ПАР УЗЛОВ СЕТИ ЦИТИРОВАНИЯ НАУЧНЫХ СТАТЕЙ

С. В. Бредихин, В. М. Ляпунов, Н. Г. Щербакова

Институт вычислительной математики и математической геофизики СО РАН
630090, Новосибирск, Россия

УДК 001.12+303.2

Представлены методы анализа сети цитирования научных статей. Определены параметры сети, отражающие свойства пар узлов: расстояние, минимальный разрез, влияние общих соседей, — и вычислены их значения. На основе данных, извлеченных из библиографической базы данных RePEc, проведен эксперимент, демонстрирующий приемы вычисления значений параметров пар и их нормирования.

Ключевые слова: библиометрия, сеть цитирования статей, расстояние, минимальный разрез, общие соседи, коэффициенты коцитирования, библиографического сочетания, ассоциативности, Адамика/Адара, Жаккара, Солтона, подобие по Кацу.

In this article methods of the analysis of the paper citation network are presented. The parameters reflecting properties of node pairs: distance, minimum cut-set, influence of common neighbors are determined and their values are calculated. The experiment showing methods of calculation of parameter values and their normalizing is fulfilled on the data retrieved from the bibliographic DB RePEc.

Key words: bibliometria, paper citation network, distance, min-cut, common neighbors, cocitation, bibliographic coupling, association, Adamic/Adar, Jaccard, Salton similarity coefficients, Katz similarity.

Введение. В данной работе продолжен анализ сети цитирования статей (СЦС), определение которой дано в [1]. Напомним: узлами ССЦ являются научные статьи, непрерывно пополняющие распределенную библиографическую базу данных (БД) RePEc [2]. Связи между узлами устанавливаются в процессе индексации информации о цитировании, содержащейся в пристатейных списках литературы. Процедура индексации содержимого БД входит в регламент ее сопровождения и периодически исполняется. В результате на момент времени T (завершение очередного процесса индексации) между цитируемой статьей i и цитирующими ее статьями устанавливаются связи, которые можно представить в виде пары $(i, \{u(i)\})$, где $\{u(i)\}$ — множество указателей на статьи, цитирующие i. Поскольку в каждый момент T к множеству существующих узлов и связей между ними происходит добавление новых узлов и связей, то СЦС можно считать продуктом процесса индексации цитирования статей в библиографической БД. Граф, представляющий СЦС, является ориентированным и ациклическим. В редких случаях ацикличность нарушается. Так, БД содержит статьи из журнала *European Journal of Political Economy* (1997), цитирующие друг друга: a) Sieg G., A model of partisan central banks and opportunistic political business cycles // European Journal of Political Economy, Volume 13, Issue 3, September 1997, Pages 503–516; b) Bergera H., Woitek U., How opportunistic are partisan German central bankers:

Evidence on the Vaubel hypothesis // European Journal of Political Economy, Volume 13, Issue 4, December 1997, Pages 807–821.

Орграф, в котором $(j,i) \in E$, если j цитирует i, будем называть *исходным*, а граф, в котором направления всех дуг поменялись на противоположные, т. е. $(i,j) \in E$, если i цитируется j, будем называть *инверсным*. Далее режим *out* будет означать, что вычисления выполнены для исходного графа, а режим *in* — для инверсного.

Для проведения численного анализа из множества всех статей, извлеченных из БД на момент времени T, были выделены те, которые одновременно цитируют хотя бы одну статью и цитируются хотя бы одной статьей. Соответствующий граф имеет одну главную компоненту слабой связности A, содержащую 131 684 вершин (97,8 % графа); компоненту A_{16}, содержащую 16 вершин, и ряд компонент с числом вершин меньше 16, не рассматриваемых в этой работе.

Центром внимания настоящей работы являются методы вычисления параметров пар узлов СЦС, основанные на измерении расстояний между узлами, анализе минимального разреза ребер и выявлении общности соседних узлов. Эти параметры характеризуют "важность" изучаемой пары. С помощью представленных далее методов можно осуществлять ранжирование и вычислить "подобие" пар узлов с учетом топологии сети. Результаты вычислений в режимах *in* и *out* представлены в виде таблиц и рисунков.

1. Параметр "расстояние между узлами". Напомним, что неориентированный граф называется *связным*, если любые две его вершины взаимно достижимы, и *несвязным* в противном случае. Ориентированный граф называется *сильно связным*, если любые две его вершины взаимно достижимы; *односторонне связным* или просто *связным*, если из каждых двух вершин по крайней мере одна достижима из другой; *слабо связным*, если при игнорировании направления ребер получается связный граф. Орграф называется *несвязным*, если он не является даже слабо связным [3]. Мы анализируем слабо связные компоненты, поэтому считаем граф связным, если он хотя бы слабо связный.

Рассмотрим неориентированный связный граф $G = (V,E)$. Для несовпадающих вершин $i, j \in V$ определим функцию $d(i,j)$, означающую длину кратчайшего пути (КП) между i и j, которая измеряется в количестве ребер. Легко проверить, что $d(i,j)$ обладает следующими свойствами:

а) $\forall\, i,j \in V$ имеет место неотрицательность $d(i,j) \geq 0$;

б) $\forall\, i,j \in V$ равенство $d(i,j) = 0$ означает, что $i = j$;

в) $\forall\, i,j \in V$ имеет место симметричность $d(i,j) = d(j,i)$;

г) $\forall\, i,j,k \in V$ имеет место неравенство треугольника $d(i,j) \leq d(i,k) + d(k,j)$.

Таким образом, на множестве вершин неориентированного связного графа $G(V,E)$ для любой пары $(i,j) \in V$ определяется структура метрического пространства, а функция $d : V \times V \to \mathbb{R}$ называется (геодезическим) *расстоянием* между вершинами i и j [4]. В случае, если граф несвязный и не существует пути из вершины i в вершину j, принято считать, что $d(i,j) = \infty$, поэтому рассматриваются обычно связные компоненты графа. Для орграфа даже в случае сильной связности может не выполняться условие (в), кроме того, принято считать, что если не существует пути из вершины i в вершину j, то такая пара не участвует при вычислении параметров. Параметр "расстояние" можно использовать для ранжирования пар вершин, т. е. пары можно упорядочить согласно неубывающим значениям длин КП между ними.

Таблица 1.1

Распределение длин КП. Режим *out*

Длина КП	Число пар A	Доля для A, %	Число пар A_{16}	Доля для A_{16}, %
1	514158	0,789982	31	63,265306
2	1467453	2,254678	16	32,653061
3	2880497	4,425760	2	6,122449
4	4217347	6,479772		
5	4979780	7,651218		
6	5134278	7,888597		
7	4965277	7,628934		
8	4777075	7,339770		
...	...	...		
27	12377	0,019017		
28	5412	0,008315		
29	2118	0,003254		
30	778	0,001195		
31	271	0,000416		
32	70	0,000108		
33	11	0,000017		
34	1	0,000002		

Эксцентриситетом $\varepsilon(i)$ вершины i в связном графе $G = (V,E)$ называется максимальное расстояние от вершины i до других вершин графа G, т. е. $\varepsilon(i) = \max\limits_{j \in V} d(i,j)$. Радиусом графа r называют минимальный эксцентриситет по всем вершинам, т. е. $r = \min\limits_{i \in V} \varepsilon(i)$. Диаметром d называют максимальный эксцентриситет по всем вершинам графа G: $d = \max\limits_{i \in V} \varepsilon(i)$ [3]. Для компоненты A значения параметров r и d таковы: $r = 1$, $d = 34$; для компоненты A_{16}: $r = 1$, $d = 3$.

В табл. 1.1 приведено распределение длин КП между вершинами компонент A и A_{16}, вычисленных в режиме *out*. На рис. 1.1 представлена гистограмма длин КП компоненты A, $F_{sp}(k)$ обозначена процентная доля КП длины k. Основой для вычислений значений параметра "расстояние" является алгоритм *BFS* [5], сложность которого оценивается как $O(|V| + |E|)$. Для A средняя длина КП равна 17. Множество КП графа (обозначим его $\{SP(G)\}$) можно построить, например, с помощью пакета *igraph* [6], или пакета *Pajek* [7]. Результаты построения $\{SP(A)\}$ для главной компоненты A с помощью вышеупомянутых программ одинаковы, $|\{SP(A)\}| = 65{,}084 \cdot 10^6$.

Пример 1. Орграф компоненты A_{16} показан на рис. 1.2, соответствие вершин с библиографическими данными — в табл. 1.2. Схема кратчайших путей описана с помощью табл. 1.3. Пусть i — строка, j — столбец. Запись $[i,j] = a$ означает: статья j цитирует статью i (путь из вершины j в вершину i имеет длину 1); $[i,j] = b$ означает: статья j цитирует некоторую статью m, которая цитирует статью i (путь из вершины j в вершину i имеет длину 2); $[i,j] = c$ означает: статья j цитирует некоторую статью m, которая, в свою очередь, цитирует статью n, цитирующую статью j (путь из вершины j в вершину i имеет длину 3); если ячейка $[i,j]$ пустая, это означает, что i и j не связаны отношением цитирования (нет пути из вершины j в i).

Для компоненты A_{16} найдены все кратчайшие пути между вершинами и вычислены их длины. Далее приведен список всех КП компоненты A_{16} (режим *out*):

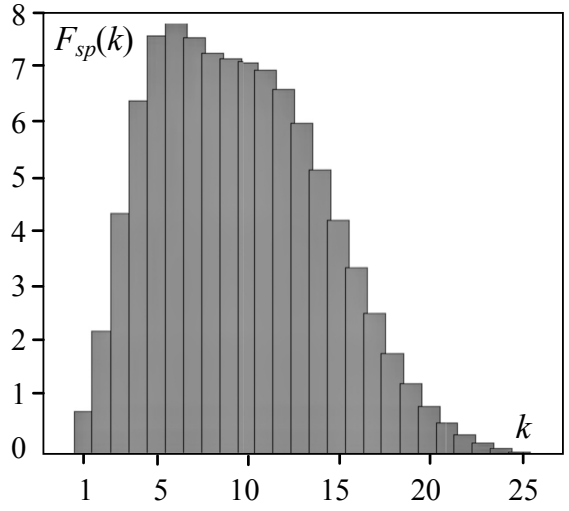

Рис. 1.1. Гистограмма длин КП компоненты A

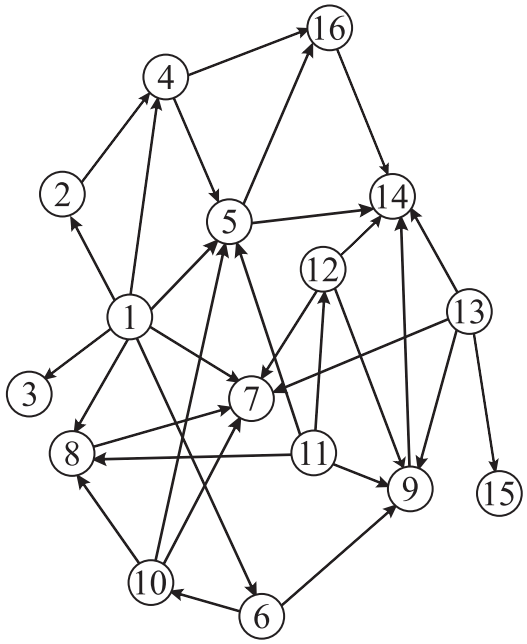

Рис. 1.2. Орграф компоненты A_{16}

$SP(A_{16})$={(1,2); (1,3); (1,4); (1,5); (1,6); (1,7); (1,8); (1,6,9); (1,6,10); (1,5,14); (1,5,16); (1,4,16);
(2,4); (2,4,5); (2,4,16,14); (2,4,5,14); (2,4,16);
(4,5); (4,16,14); (4,5,14); (4,16);
(5,14); (5,16);
(6,10,5); (6,10,7); (6,10,8), (6,9); (6,10); (6,9,14); (6,10,5,16);
(8,7);
(9,14);
(10,5); (10,7); (10,8); (10,5,14); (10,5,16);
(11,5); (11,12,7); (11,8,7); (11,8); (11,9); (11,12); (11,12,14); (11,9,14);
(11,5,14); (11,5,16);
(12,7); (12,9); (12,14);
(13,7); (13,9); (13,14); (13,15);
(16,14)}.

2. Параметр "минимальный разрез". Термин "*cut*" применительно к графу означает "разрез" ребра (или ребер), в результате которого вершины разделяются на два непересекающихся подмножества так, что смежные вершины относительно разрезанных ребер попадают в разные подмножества. Рассмотрим граф $G = (V,E)$ и две несмежные вершины i, $j \in V$, такие что существует путь из i в j. Вопрос: какова "сила связи" между i и j? Ответ на него можно сформулировать так: какое минимальное количестве ребер необходимо разрезать, чтобы потерялась связь между i и j.

Задача о разрезе графа является обобщением задачи о потоке в транспортной сети (*flow network*), представленой орграфом $G = (V,E)$ и функцией $u : E \to \mathbb{R}$, приписывающей каждому ребру $e = (i,j) \in E$ пропускную способность $u(e) \geq 0$. Для двух различных вершин $s,t \in V$ (источник и приемник) функция $f : E \to \mathbb{R}$ называется потоком от s к t (или (s-t)-потоком), если она удовлетворяет условиям a) емкости и b) баланса. Формально:

a) $\forall e \in E : 0 \le f(e) \le u(e)$;

b) $\forall v \in V \setminus \{s,t\} : \sum\limits_{e \in \Gamma^-(v)} f(e) = \sum\limits_{e \in \Gamma^+(v)} f(e)$.

Таблица 1.2

Связь вершин орграфа компоненты A_{16} с идентификаторами статей в БД и их библиографическими данными

Статья i	Идент. i в БД	Библиографические данные i
1	55774	Mishra, S.K.; Lai, K.K. Second order symmetric duality in multiobjective programming involving generalized cone-invex functions // European J. of Operational Research, 2007, vol.178, issue 1, pp. 20-26
2	55775	Yang, X. M.; Yang, X. Q.; Teo, K. L.; Hou, S. H. Second order symmetric duality in non-differentiable multiobjective programming with F-convexity // European J. of Operational Research, 2005, vol.164, issue 2, pp. 406-416
3	58606	Yang, X. M.; Yang, X. Q.; Teo, K. L.; Hou, S. H. Multiobjective second-order symmetric duality with F-convexity // European J. of Operational Research, 2005, vol.165, issue 3, pp. 585-591
4	78389	Yang, X. M.; Yang, X. Q.; Teo, K. L. Non-differentiable second order symmetric duality in mathematical programming with F-convexity // European J. of Operational Research, 2003, vol.144, issue 3, pp. 554-559
5	78390	Mishra, S. K. Second order symmetric duality in mathematical programming with F-convexity // European J. of Operational Research, 2000, vol.127, issue 3, pp. 507-518
6	78241	Khurana, Seema Symmetric duality in multiobjective programming involving generalized cone-invex functions // European J. of Operational Research, 2005, vol.165, issue 3, pp. 592-597
7	117825	Kim, Do Sang; Yun, Ye Boon; Lee, Won Jung. Multiobjective symmetric duality with cone constraints // European J. of Operational Research, 1998, vol.107, issue 3, pp. 686-691
8	117824	Mishra, S. K. Multiobjective second order symmetric duality with cone constraints // European J. of Operational Research, 2000, vol.126, issue 3, pp. 675-682
9	78245	Chandra, S.; Kumar, V. A note on pseudo-invexity and symmetric duality // European J. of Operational Research, 1998, vol.105, issue 3, pp. 626-629
10	106436	Suneja, S. K.; Aggarwal, Sunila; Davar, Sonia Multiobjective symmetric duality involving cones // European J. of Operational Research, 2002, vol.141, issue 3, pp. 471-479
11	78242	Chen, Xiuhong. Minimax and symmetric duality for a class of multiobjective variational mixed integer programming problems // European J. of Operational Research, 2004, vol.154, issue 1, pp. 71-83
12	78244	Kim, Do Sang; Song, Young Ran. Minimax and symmetric duality for nonlinear multiobjective mixed integer programming // European J. of Operational Research, 2001, vol.128, issue 2, pp. 435-446.
13	78243	Chandra, Suresh; Abha. Technical note on symmetric duality in multiobjective programming: Some remarks on recent results // European J. of Operational Research, 2000, vol.124, issue 3, pp. 651-654
14	82175	Kumar, V.; Husain, I.; Chandra, S. Symmetric duality for minimax nonlinear mixed integer programming // European J. of Operational Research, 1995, vol.80, issue 2, pp. 425-430
15	131949	Das, L. N.; Nanda, S. Symmetric dual multiobjective programming // European J. of Operational Research, 1997, vol.97, issue 1, pp. 167-171
16	78391	Gulati, T. R.; Ahmad, Izhar. Second order symmetric duality for nonlinear minimax mixed integer programs // European J. of Operational Research, 1997, vol.101, issue 1, pp. 122-129

Таблица 1.3

Схема кратчайших путей компоненты A_{16}

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1																
2	a															
3	a															
4	a	a														
5	a	b		a		b				a	a					
6	a															
7	a					b		a		a	b	a	a			
8	a					b				a	a					
9	b					a					a	a	a			
10	b					a										
11																
12											a					
13																
14	b	c		b	a	b			a	b	b	a	a			a
15													a			
16	b	b		a	a	c				b	b					

Величина потока f определяется путем вычисления разности

$$\sum_{e\in\Gamma^+(s)} f(e) - \sum_{e\in\Gamma^-(s)} f(e),$$

где $\Gamma^+(v)$ — множество ребер, исходящих из v; $\Gamma^-(v)$ — множество ребер, входящих в v [8]. Задача о максимальном потоке (*max-flow problem*) состоит в нахождении потока максимальной величины.

Для графа $G = (V,E)$ *разрез* (*cut*) означает разбиение вершин графа на два непересекающихся подмножества $C = (S,T)$: $S \cup T = V$ и $S \cap T = \varnothing$. Каждый разрез определяет *множество ребер разреза* (*cut-set*), такое что одна вершина ребра принадлежит подмножеству S, а вторая — T: *cut-set*$(C) = \{(i,j) \in E | i \in S, j \in T\}$. В этом случае говорят, что ребро (i,j) пересекает разрез. Если граф ориентированный, то ребра, пересекающие разрез, ориентированные и направлены из S в T. В невзвешенном графе *размер разреза* — это количество пересекающих его ребер. Во взвешенном графе *величина разреза* — это сумма весов пересекающих его ребер [5]. Если в графе выделены две вершины s и t, то (s-t)-*разрез* — это $C = (S,T), T = V - S (s \in S, t \in T)$ [9]. Найти *минимальный* (s-t)-разрез значит определить такие S и T, чтобы величина разреза была минимальной. Отметим, что термин *разрез* в зависимости от контекста означает как разбиение множества вершин, так и множества ребер. Из теоремы Форда — Фалкерсона [10] следует, что для транспортной сети максимальная величина (s-t)-потока равна величине минимального (s-t)-разреза.

В этом контексте будем рассматривать "силу связи" между двумя статьями (вершинами s и t) как задачу о минимальном (s-t)-разрезе. Обозначим через Q множество всех пар вершин (s,t) компоненты A, таких что t достижима из s; $|Q| = 514\,158$. Отметим, что число всех пар A оценивается как $8{\times}10^9$. Множество Q было получено на этапе вычисления КП. Для всех пар из множества Q вычислены размеры минимальных разрезов. Гистограмма размеров (s-t)-разрезов с логарифмическим масштабом по оси ординат представлена на рис. 2.1, из которого видно, что для значительной части пар (64,78 %) достаточно разреза одного ребра, чтобы вершины разделились на два множества, S и T. Вычисление размеров (s-t)-разрезов для компоненты A и нахождение ребер всех минимальных разрезов для

каждой пары s, t компоненты A_{16} выполнены с помощью пакета *igraph*. Сложность алгоритма, основанного на вычислении величины максимального потока, предложенного в работе [11], составляет $\mathcal{O}(|V|^3)$. Сложность алгоритма нахождения ребер всех минимальных разрезов, предложенного в работе [9], составляет $\mathcal{O}(n(|V| + |E|)) + \mathcal{O}(F)$, где n — количество разрезов, а $\mathcal{O}(F)$ — сложность алгоритма вычисления максимального потока.

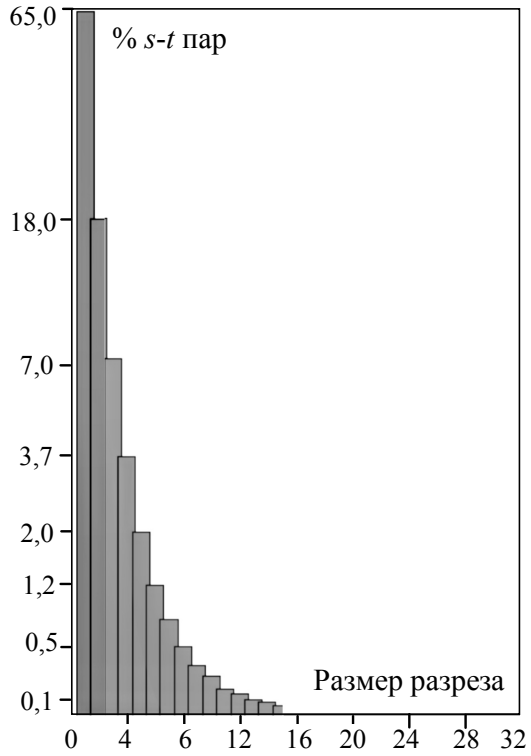

Рис. 2.1. Гистограмма размеров s-t разрезов. Компонента A

Чтобы сформулировать условия для алгоритма порождения всех (s-t)-разрезов, введем дополнительные определения. Пусть имеются множества $X \subseteq V$ и $Y = V - X$. Обозначим через $\Gamma(X)$ множество внешних соседей вершин множества X: $\Gamma(X) = \{v \in Y \ \& \ \exists\, u \in X : (u,v) \in E\}$. Множеству X соответствует подграф $< X >= (X, E_X)$, такой что $E_X = \{(u,v) \in E : u,v \in X\}$. Пусть $u,v \in X$, вершина v считается достижимой из u в $< X >$ (обозначается $u[X]v$), если либо $u = v$, либо существует путь из u в v в $< X >$.

Для того чтобы набор ребер C был минимальным (s-t)-разрезом орграфа $G = (V,E)$, необходимо и достаточно, чтобы он мог быть определен как $C = (S,T)$, $T = V - S$, где S должно удовлетворять трем условиям [9]:

$$a)\ s \in S,\ \ t \notin S;$$
$$b)\ \forall x \in S\ \ s[S]x;$$
$$c)\ \forall y \in \Gamma(S)\ \ y[T]t.$$

Понятие разреза используется для решения задач кластеризации. Примеры алгоритмов кластеризации на основе минимального разреза можно найти в работах [12–14]. Идея заключается в создании кластеров с достаточно малыми значениями разрезов для межкластерных связей и достаточно большими значениями для внутрикластерных связей.

Пример 2. Для компоненты A_{16} для всех пар вершин $(s - t)$ в режиме *out* вычислены размеры минимальных (s-t)-разрезов и определены соответствующие наборы ребер. Заметим, что в результате не только не существует путей из s в t, но и путей из вершин, относящихся к множеству S, в вершины множества T. В общем случае обратное не гарантируется. Из 49 пар $(s - t)$, таких что t достижима из s, для 33 (67,35 %) достаточно разрезать одно ребро, чтобы вершины графа разделились на два множества S и T, таких что из вершин множества S нет путей в T; для 12 пар (24,49 %) достаточно разрезать два ребра; для 4 пар достаточно разрезать три ребра. Минимальных разрезов для пары $(s-t)$ может быть несколько. В табл. 2.1 приведены варианты минимальных разрезов для пары вершин (11–7). В колонке “Ребра размера” указаны наборы разрезаемых ребер; в колонке S приведен состав вершин этого множества. Рис. 2.2, на котором выделены разрезаемые ребра и множество S, иллюстрирует эти варианты.

3. Параметры, зависящие от “общих соседей”. Рассмотрим параметры подобия вершин графа $G = (V, E)$, основанные на общности множества ближайших соседей

Таблица 2.1
Варианты минимальных (s-t)-разрезов пары (11-7) компоненты A_{16}

$s-t$	Ребра разреза	Множество S	Рисунок
11–7	(8,7); (11,12)	S={5, 8, 9, 11, 14, 16}	Рис. 2.2,*а*
11–7	(8,7); (12,7)	S={5, 8, 9, 11, 12, 14, 16}	Рис. 2.2,*б*

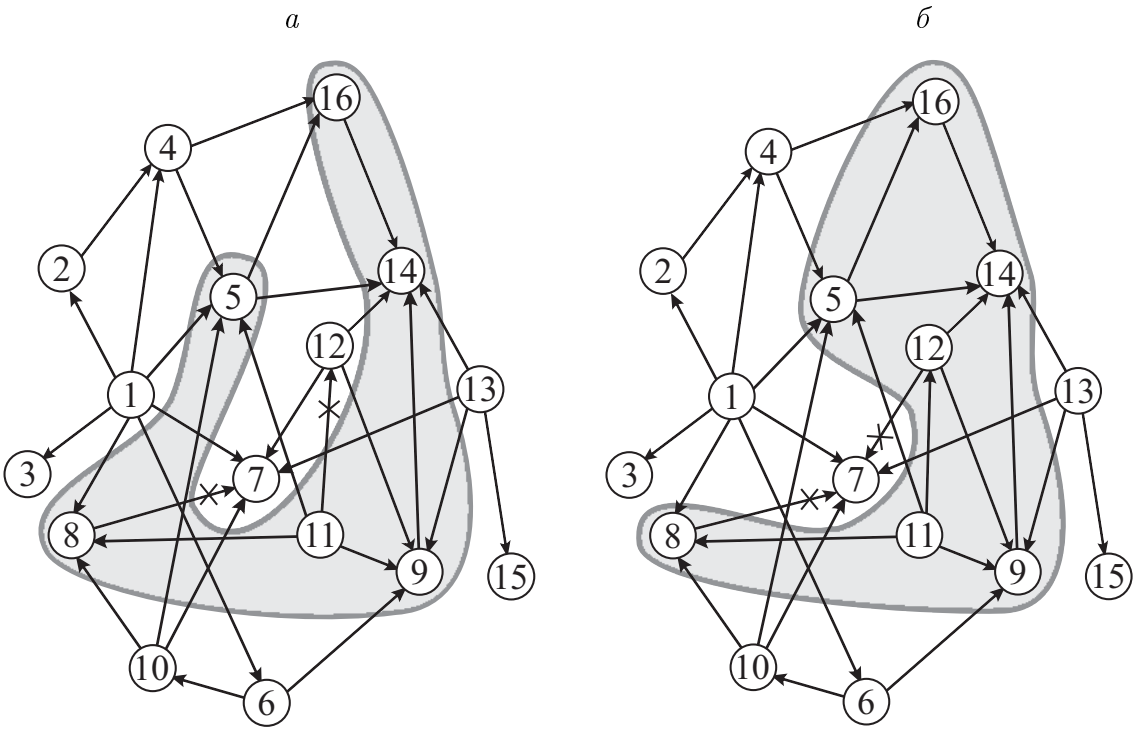

Рис. 2.2. (11-7)-разрез компоненты A_{16}: *а* — вариант а, *б* — вариант б

(*common nearest neighbors*). Условимся обозначать ребро между вершинами i и j неориентированного графа как $[i,j]$, а ребро орграфа, направленное из i в j, как (i,j). Обозначим через $\Gamma(i)$ множество ближайших соседей вершины i. Для неориентированного графа ближайшими соседями вершины i являются все вершины, соединенные с i ребром: $\Gamma(i) = \{j : [i,j] \in E\}$. Для орграфа, согласно определению, приведенному в монографии [5], ближайшими соседями вершины i также являются все вершины, соединенные с ней ребром, независимо от направления: $\Gamma(i) = \{j : (i,j) \in E \vee (j,i) \in E\}$. В данной работе ближайшие соседи вершины i определяются с учетом ориентации: $\Gamma^+(i) = \{j : (i,j) \in E\}$, $\Gamma^-(i) = \{j : (j,i) \in E\}$. Мы рассматриваем $\Gamma^+(i)$ для исходного и инверсного графов, поэтому будем использовать $\Gamma(i)$. Рис. 3.1 иллюстрирует понятие общих соседей для орграфа, здесь вершины d, e являются ближайшими общими соседями вершин a и b.

На параметрах подобия пар сетевых узлов строятся алгоритмы извлечения информации, кластеризации вершин сети, предсказания путей эволюции сетей, например, вероятность появления новых связей между вершинами (*link prediction*) [15, 16]. В библиометрии подобие сетевых акторов используется при поиске и извлечении информации из БД [17, 18], выявлении множества ключевых статей и тематик [19, 20], построении визуального представления областей знаний и создания карт науки [21–23].

3.1. *Параметр CN* определяет степень подобия вершин графа исходя из количества ближайших общих соседей, вычисляется по формуле:

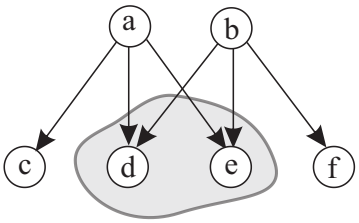

Рис. 3.1. Общие соседи

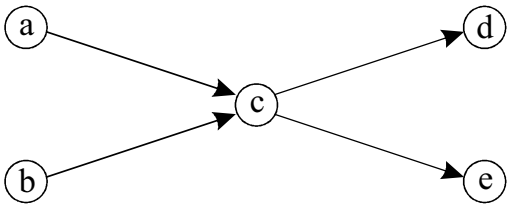

Рис. 3.2. Отношения библиографического сочетания и коцитирования

$$CN(i,j) = |\Gamma(i) \cap \Gamma(j)|.$$

Подход изложен в работе [24], где подобие рассматривается применительно к сети соавторства (ребро $[i,j] \in E$ тогда и только тогда, когда авторы i и j являются соавторами одной и более работ). Изучалась корреляция между количеством общих соседей у авторов i и j ко времени t и вероятностью того, что они будут сотрудничать в будущем (появится ребро $[i,j]$ ко времени $T > t$). К подобию, строящемуся на рассмотрении ближайших общих соседей, применяется термин *структурная эквивалентность* (*structural equivalence*). При определении *регулярной эквивалентности* учитывается, насколько соседи вершин сами являются подобными. В этом случае коэффициент подобия σ_{ij} для пары вершин i и j можно представить в виде [25]:

$$\sigma_{ij} = \alpha \sum_{kl} c_{ik} c_{jl} \sigma_{kl}.$$

В СЦС ближайшими общими соседями вершин i,j является множество статей, на которые ссылаются обе статьи. В библиометрии $CN(i,j)$ — это *коэффициент библиографического сочетания* (*КБС*). Научные статьи связаны отношением библиографического сочетания, если они одновременно цитируют хотя бы одну статью [26]. Рис. 3.2 поясняет это определение: статьи a и b находятся в отношении библиографического сочетания по отношению к статье c. В терминах матрицы смежности $C = [c_{ij}]$, $c_{ij} = 1$, если ребро $(j,i) \in E$ (т. е. статья j цитирует статью i), *КБС* можно вычислить по формуле

$$\text{КБС}(i,j) = \sum_{k=1}^{n} c_{ki} c_{kj} = \left(C^\top C\right)_{ij}. \tag{3.1}$$

Здесь и далее n — количество вершин графа.

Рассмотрим инверсный граф (режим *in*) В соответствующей матрице смежности $c_{ij} = 1$, если ребро $(i,j) \in E$. Тогда формула (3.1) будет определять *коэффициент коцитирования* (*ККЦ*); статьи считаются связанными отношением коцитирования, если их одновременно цитирует хотя бы одна статья [27, 28]. Рис. 3.2 иллюстрирует это определение; статьи d и e находятся в отношении коцитирования по отношению к статье c. На практике *ККЦ* можно вычислить из матрицы исходного графа:

$$\text{ККЦ}(i,j) = \sum_{k=1}^{n} c_{ik} c_{jk} = \left(CC^\top\right)_{ij}. \tag{3.2}$$

Поскольку значение коэффициента *CN* существенно зависит от степени вершины, например, от принятых в разных сообществах норм цитирования, популярности авторов

статьи и т. д., коэффициенты принято нормировать. Нормированные значения, изменяющиеся в диапазоне от 0 до 1, называют мерой подобия. Рассмотрим ряд способов нормирования на примере коцитирования.

3.2. *Нормирование по Солтону.* Пусть $cit(i)$ — количество цитирований, полученных статьей i; $coc(i,j)$ — *коэффициент коцитирования* (подобия); S — нормированное значение подобия. Нормирование по Солтону (*косинусное подобие*) [29] имеет вид:

$$S_S(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{(|\Gamma(i)| \times |\Gamma(j)|)^{\frac{1}{2}}} = \frac{coc(i,j)}{(cit(i) \times cit(j))^{\frac{1}{2}}}, \tag{3.3}$$

т. е. вычисляется отношение между количеством общих цитирований и геометрическим средним количеств цитирований каждой статьи. Нормирование по Солтону относится к числу наиболее часто используемых в библиометрии.

3.3. *Нормирование по Жаккару* (*индекс Жаккара*) в применении к цитированиям выглядит следующим образом:

$$S_J(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} = \frac{coc(i,j)}{cit(i) + cit(j) - coc(i,j)}, \tag{3.4}$$

т. е. вычисляется отношение между количеством общих цитирований и количеством цитирований хотя бы одной из рассматриваемой пары статей.

3.4. *Коэффициент ассоциативности* определен в работе [30] (то же, что *proximity index* [31]); формула для отношения коцитирования имеет вид:

$$S_A(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i)| \times |\Gamma(j)|} = \frac{coc(i,j)}{cit(i) \times cit(j)}. \tag{3.5}$$

Коэффициент вычисляется как отношение между количеством общих цитирований и ожидаемым количеством общих цитирований в предположении, что цитирования i и j статистически независимы. В работе [32] проведен анализ часто используемых методов нормирования и сделан вывод в пользу коэффициента ассоциативности.

3.5. *Коэффициент Adamic/Adar.* В работе [33] выдвинуто предположение, что свойство подобия выражено сильнее для вершин, имеющих больше общих соседей, при том что эти соседи редко являются соседями для других вершин, так как в противном случае общность может оказаться случайной. Для неориентированного графа коэффициент определяется согласно формуле

$$S_{AA}(i,j) = \sum_{z \in (\Gamma(i) \cap \Gamma(j))} \frac{1}{\ln(deg(z))},$$

где $deg(z)$ — степень вершины z. Для случая библиографического сочетания коэффициент определяется как

$$S_{AA}(i,j) = \sum_{z \in (\Gamma(i) \cap \Gamma(j))} \frac{1}{\ln(cit(z))}. \tag{3.6}$$

Здесь z — публикация, которую цитируют публикации i и j, $cit(z)$ — насколько часто цитируют z. В случае определения коэффициента коцитирования z — публикация, которая

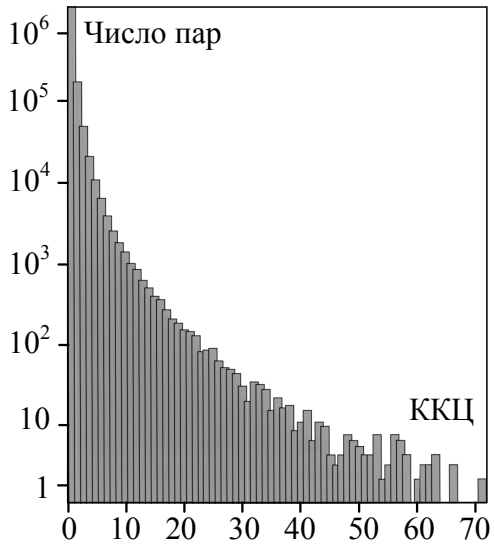

Рис. 3.3. Гистограмма *ККЦ* компоненты A

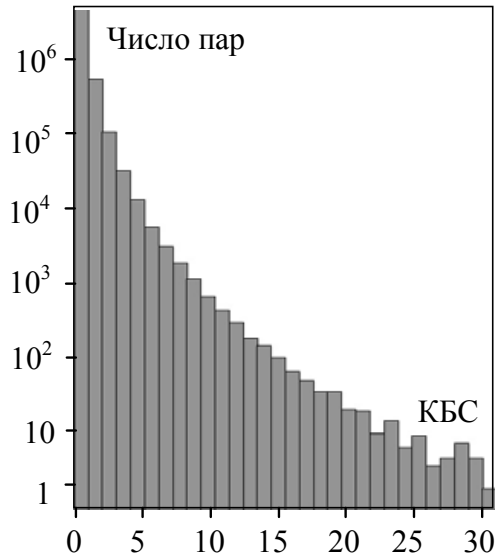

Рис. 3.4. Гистограмма *КБС* компоненты A

цитирует публикации i и j, а в знаменателе появляется логарифм количества публикаций, цитируемых публикацией z. Данный коэффициент чаще используется для случая неориентированных графов и реже для анализа подобия в библиометрии.

3.6. *Вычислительный эксперимент.* Из (3.1) и (3.2) следует, что коэффициенты подобия можно получить путем умножения матриц. В нашем случае матрица смежности C имеет порядок 131 684. Поскольку результирующая матрица $(CC^\top)$ является разреженной, для вычисления коэффициентов подобия было использовано представление данных о цитировании в виде списков с элементами $(i,\{u(i)\})$.

Компонента A имеет $8{,}7\times10^8$ пар статей, из которых $1{,}7\times10^6$ пар (0,02 %) связаны отношением коцитирования. Исследование показало, что для компоненты A $ККЦ = 1$ для 85,7 % пар коцитируемых статей, и только для 8,9 % пар коэффициент равен двум. Максимальное значение $ККЦ = 399$ имеет одна пара статей (7619, 7622, табл. 3.1).

Отношением библиографического сочетания связаны $6{,}1\times10^6$ пар (0,07 %), из них 86,6 % пар имеют коэффициент *КБС*, равный единице, а 10,2 % пар имеют коэффициент, равный двум. Максимальное значение коэффициента *КБС*, равное 43, имеет одна пара (7220, 6511, см. табл. 3.1). Гистограммы коэффициентов библиографического сочетания и коцитирования представлены на рис. 3.3, 3.4. По оси ординат указано количество пар в логарифмическом масштабе, а по оси абсцисс — соответствующий коэффициент подобия.

Для статей компоненты A вычислены значения коэффициентов *ККЦ*, *КБС*, S_J, S_A и S_{AA}. В табл. 3.1 приведены результаты, ранжированные согласно значениям ненормированных коэффициентов. Табл. 3.2 устанавливает соответствие между идентификаторами статей из табл. 3.1 и их библиографическими данными. Для выявления, насколько различные способы вычисления линейно связаны между собой, вычислены соответствующие коэффициенты Пирсона r, монотонную зависимость отражают коэффициенты ранговой корреляции Спирмена ρ. Результаты сравнения приведены в табл. 3.3. Коэффициенты r и ρ выявили существенную разницу между ранжированием согласно ненормированным значениям и S_A.

Таблица 3.1

Коэффициенты подобия пар компоненты A (первые 12 позиций)

Режим *in*						Режим *out*					
Ранг по *ККЦ*	Идентификаторы статей в БД	*ККЦ* (3.2)	S_J (3.4)	S_A (3.5)	S_{AA} (3.6)	Ранг по *КБС*	Идентификаторы статей в БД	*КБС* (3.1)	S_J (3.4)	S_A (3.5)	S_{AA} (3.6)
1	7619, 7622	399	0,3950	0,0008	189,0950	1	7220, 6511	43	0,4257	0,0083	11,5040
2	25704, 71824	340	0,4480	0,0011	160,2397	2	7220, 7252	36	0,3711	0,0083	8,9622
3	7619, 32100	176	0,2071	0,0010	74,6000	3–4	13979, 13985	34	0,9189	0,0270	14,2395
4	32100, 7622	126	0,1850	0,0010	51,9201	3–4	7225, 7234	34	0,3820	0,0090	10,9527
5	25704, 25774	104	0,1559	0,0011	43,6626	5–6	7220, 6516	32	0,3368	0,0082	8,4120
6	51987, 95510	101	0,0991	0,0003	74,7790	5–6	7220, 15245	32	0,3765	0,0102	8,5922
7	86325, 86323	100	0,2667	0,0018	49,0610	7	20345, 20343	30	0,9091	0,0303	12,0823
8	117075, 125984	99	0,8115	0,0082	54,0653	8–11	15407, 15402	29	0,6042	0,0203	7,7429
9	65149, 108187	98	0,2620	0,0026	36,8778	8–11	11342, 11343	29	0,8056	0,0278	0,1905
10	25704, 38664	93	0,1380	0,0010	38,9616	8–11	7220, 15402	29	0,3152	0,0084	7,6914
11	25774, 71824	91	0,1707	0,0013	38,2412	8–11	7220, 24374	29	0,3766	0,0125	7,3825
12	32380, 62029	84	0,2667	0,0021	40,5752	12	7240, 2306	28	0,5833	0,0194	7,0584

Таблица 3.2

Связь между идентификаторами статей и их библиографическими данными для результатов, приведенных в табл. 3.1

Идент. i в БД	Библиографические данные i
2306	Kanjilal, Kakali; Ghosh, Sajal. Environmental Kuznet's curve for India: Evidence from tests for cointegration with unknown structuralbreaks // Energy Policy, 2013, vol. 56, issue C, pp. 509-515.
6511	Muhammad Shahbaz; Mete Feridun. Electricity consumption and economic growth empirical evidence from Pakistan Quality & Quantity // International J. of Methodology, 2012, vol. 46, issue 5, pp. 1583-1599
6516	Muhammad Shahbaz; Mete Feridun. Electricity consumption and economic growth empirical evidence from Pakistan // Quality & Quantity: International J. of Methodology, 2012, vol. 46, issue 5, pp. 1583-1599
7220	Ozturk, Ilhan. A literature survey on energy-growth nexus // Energy Policy, 2010, vol. 38, issue 1, pp. 340-349
7225	Saboori, Behnaz; Sulaiman, Jamalludin. Environmental degradation, economic growth and energy consumption: Evidence of the environmental Kuznets curve in Malaysia // Energy Policy, 2013, vol. 60, issue C, pp. 892-905
7234	Al-mulali, Usama; Fereidouni, Hassan Gholipour; Lee, Janice Ym; Sab, Che Normee Binti Che. Examining the bi-directional long run relationship between renewable energy consumption and GDP growth // Renewable and Sustainable Energy Reviews, 2013, vol. 22, issue C, pp. 209-222
7240	Ghosh, Sajal. Examining carbon emissions economic growth nexus for India: A multivariate cointegration approach // Energy Policy, 2010, vol. 38, issue 6, pp. 3008-3014
7252	Jalil, Abdul. Energy-growth conundrum in energy exporting and importing countries: Evidence from heterogeneous panel methods robust to cross-sectional dependence // Energy Economics, 2014, vol. 44, issue C, pp. 314-324
7619	Blundell, Richard; Bond, Stephen. Initial conditions and moment restrictions in dynamic panel data models // J. of Econometrics, 1998, vol. 87, issue 1, pp. 115-143
7622	Arellano, Manuel; Bover, Olympia. Another look at the instrumental variable estimation of error-components models // J. of Econometrics, 1995, vol. 68, issue 1, pp. 29-51
11342	Sueyoshi, Toshiyuki; Goto, Mika. DEA environmental assessment in a time horizon: Malmquist index on fuel mix, electricity and CO2 of industrial nations // Energy Economics, 2013, vol. 40, issue C, pp. 370-382
11343	Sueyoshi, Toshiyuki; Goto, Mika; Sugiyama, Manabu. DEA window analysis for environmental assessment in a dynamic time shift: Performance assessment of U.S. coal-fired power plants // Energy Economics, 2013, vol. 40, issue C, pp. 845-857
13979	Wang, Lanfang; Wang, Susheng. Economic freedom and cross-border venture capital performance // J. of Empirical Finance, 2012, vol. 19, issue 1, pp. 26-50
13985	Chortareas, Georgios E.; Girardone, Claudia; Ventouri, Alexia. Financial freedom and bank efficiency: Evidence from the European Union // J. of Banking & Finance, 2013, vol. 37, issue 4, pp. 1223-1231
15245	Kouakou, Auguste K. Economic growth and electricity consumption in Cote d'Ivoire: Evidence from time series analysis // Energy Policy, 2011, vol. 39, issue 6, pp. 3638-3644
15402	Apergis, Nicholas; Payne, James E. A dynamic panel study of economic development and the electricity consumption-growth nexus // Energy Economics, 2011, vol. 33, issue 5, pp. 770-781
15407	Payne, James E. A survey of the electricity consumption-growth literature // Applied Energy, 2010, vol. 87, issue 3, pp. 723-731
20343	Ou, Xunmin; Xiaoyu, Yan; Zhang, Xiliang. Life-cycle energy consumption and greenhouse gas emissions for electricity generation and supply in China // Applied Energy, 2011, vol. 88, issue 1, pp. 289-297

Продолжение таблицы 3.2

1	2
20345	Sharma, Susan Sunila. Determinants of carbon dioxide emissions: Empirical evidence from 69 countries // Applied Energy, 2011, vol. 88, issue 1, pp. 376-382
24374	Chang, Ching-Chih; Soruco Carballo, Claudia Fabiola. Energy conservation and sustainable economic growth: The case of Latin America and the Caribbean // Energy Policy, 2011, vol. 39, issue 7, pp. 4215-4221
25704	Im, Kyung So; Pesaran, M. Hashem; Shin, Yongcheol. Testing for unit roots in heterogeneous panels // J. of Econometrics, 2003, vol. 115, issue 1, pp. 53-74
25774	Choi, In. Unit root tests for panel data // J. of International Money and Finance, 2001, vol. 20, issue 2, pp. 249-272
32100	Windmeijer, Frank. A finite sample correction for the variance of linear efficient two-step GMM estimators // J. of Econometrics, 2005, vol. 126, issue 1, pp. 25-51
32380	Berger, Allen N.; Humphrey, David B. Efficiency of financial institutions: International survey and directions for future research // European J. of Operational Research, 1997, vol. 98, issue 2, pp. 175-212
38664	M. Hashem Pesaran. A simple panel unit root test in the presence of cross-section dependence // J. of Applied Econometrics, 2007, vol. 22, issue 2, pp. 265-312
51987	Engle, Robert F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation // Econometrica, 1982, vol. 50, issue 4, pp. 987-1007
62029	Berger, Allen N.; Mester, Loretta J. Inside the black box: What explains differences in the efficiencies of financial institutions? // J. of Banking & Finance, 1997, vol. 21, issue 7, pp. 895-947
65149	Paresh Kumar Narayan. The saving and investment nexus for China: evidence from cointegration tests // Applied Economics, 2005, vol. 37, issue 17, pp. 1979-1990
71824	Levin, Andrew; Lin, Chien-Fu; James Chu, Chia-Shang. Unit root tests in panel data: asymptotic and finite-sample properties // J. of Econometrics, 2002, vol. 108, issue 1, pp. 1-24
86323	Pesaran, H. Hashem; Shin, Yongcheol. Generalized impulse response analysis in linear multivariate models // Economics Letters, 1998, vol. 58, issue 1, pp. 17-29
86325	Koop, Gary; Pesaran, M. Hashem; Potter, Simon M. Impulse response analysis in nonlinear multivariate models // J. of Econometrics, 1996, vol. 74, issue 1, pp. 119-147
95510	Bollerslev, Tim; Chou, Ray Y.; Kroner, Kenneth F. ARCH modeling in finance : A review of the theory and empirical evidence // J. of Econometrics, 1992, vol. 52, issue 1-2, pp. 5-59
108187	M. Hashem Pesaran; Yongcheol Shin; Richard J. Smith. Bounds testing approaches to the analysis of level relationships // J. of Applied Econometrics, 2001, vol. 16, issue 3, pp. 289-326
117075	Margaret M. McConnell; Gabriel Perez-Quiros. Output fluctuations in the United States: what has changed since the early 1980s? // Proceedings, 2000, issue Mar
125984	Gabriel Perez-Quiros; Margaret M. McConnell. Output Fluctuations in the United States: What Has Changed since the Early 1980's? // American Economic Review, 2000, vol. 90, issue 5, pp. 1464-1476

Чтобы определить, насколько каждый способ вычисления параметра подобия зависит от степени вершины (*indeg* для режима *in*, *outdeg* для режима *out*) вычислены зависимости (r, ρ) между степенью вершины и средним значением меры подобия с другими вершинами для всех способов вычисления меры. Например, для того, чтобы определить, насколько параметр *КБС* зависит от размера, выполняются два ранжирования: 1) вычисляется ранг вершины i согласно значению $outdeg(i)$; 2) вычисляется ранг вершины i согласно среднему значению коэффициента подобия с остальными вершинами:

$$\sum_{j:\Gamma(i)\cap\Gamma(j)\neq\varnothing} \text{КБС}(i,j) \Big/ \big|\{j : \Gamma(i)\cap\Gamma(j)\neq\varnothing\}\big|.$$

Таблица 3.3

Корреляция коэффициентов подобия

	$ККЦ - S_J$	$ККЦ - S_A$	$ККЦ - S_{AA}$	$КБС - S_J$	$КБС\ - S_A$	$КБС - S_{AA}$
r	0,048422	−0,065360	0,929088	0,158967	−0,590200	0,768028
ρ	0,379987	0,250849	0,667335	0,479550	0,317241	0,636459

Таблица 3.4

Влияние размера на значения коэффициентов подобия

Входящая и исходящая степени	Ненормированный коэффициент	S_J (3.4)	S_A (3.5)	S_{AA} (3.6)
Indeg	r: 0,370150	r : −0,249915	r: −0,243130	r: 0,078785
	ρ: 0,730420	ρ: −0,596163	ρ: −0,755543	ρ: 0,188294
Outdeg	r: 0,447588	r: −0,517798	r : −0,492412	r: −0,096400
	ρ: 0,692119	ρ: −0,873253	ρ: −0,900179	ρ: −0,080632

Для полученных множеств вычисляются коэффициенты r и ρ. Результаты приведены в табл. 3.4. Для компоненты A ненормированный способ вычисления сильнее других зависит от "размера", а S_A меньше других. Однако независимость S_A от размера может не позволить использовать его для установления отношения подобия среди публикаций, относящихся к одной области знаний, а значит, в среднем, пользующихся одними и теми же нормами цитирования.

Пример 3. Результаты вычисления коэффициентов подобия для компоненты A_{16} приведены в табл. 3.5. Множество пар можно разделить на три группы согласно значениям *ККЦ*, на пять групп согласно значениям коэффициента S_{AA}, на 9 групп согласно S_J и на 12 групп согласно S_A.

Приведем свойства нормированных коэффициентов подобия.

Свойство 3.1. Из равенства $CN(i,j) = deg(i) = deg(j)$ следует, что $S_J(i,j)$ принимает максимальное значение. Здесь и далее $deg(i)$ в случае орграфа обозначает соответствующую степень вершины в исходном графе (*outdeg* для библиографического сочетания и *indeg* для коцитирования). Так, пары статей i, j, у которых количество цитирований и коцитирований совпадает $ККЦ(i,j) = indeg(i) = indeg(j)$, имеют наибольшее значение S_J. Для коэффициента S_A это верно только для случая, когда все составляющие равны единице. Среди таких пар статей большую часть составляют статьи, имеющие по одному цитированию. Они наиболее сходны между собой, но наименее интересны, поскольку при представлении результатов обычно применяются пороги, ограничивающие множество объектов.

Свойство 3.2. Если $S_J(i,j)$ или $S_A(i,j)$ принимает максимальное значение, то выполняются равенства $CN(i,j) = deg(i) = deg(j)$.

Свойство 3.3. Если $CN(i_1,j_1) = CN(i_2,j_2)$ и $\{deg(i_1), deg(j_1)\} = \{deg(i_2), deg(j_2)\}$, то $S_J(i_1,j_1) = S_J(i_2,j_2)$ и $S_A(i_1,j_1) = S_A(i_2,j_2)$. Т. е. если две пары имеют одинаковое число общих соседей и множества степеней для пар одинаковы, то значения коэффициентов S_J/S_A одинаковы.

Для коэффициента S_{AA} на компоненте A_{16} построен пример, демонстрирующий, что свойства 3.1–3.3 для этого коэффициента не выполняются.

4. Параметр "подобие по Кацу". В работе [34] определяется статус вершины неориентированного графа, зависящий от статуса связанных с ней вершин. Сила взаимосвязи между вершинами определена как функция от длины путей между вершинами и вычисляется по формуле

$$S_{Katz}(i,j) = \beta^l \sum_{l=1}^{\infty} |paths_{i,j}^{<l>}|, \tag{4.1}$$

где $paths_{i,j}^{<l>}$ = {путь длиной l от вершины i до вершины j}; $0 \le \beta \le 1$ — параметр затухания. $S_{Katz}(i,j)$ принято называть коэффициентом подобия пары вершин по Кацу. Из (4.1) следует, что короткие пути вносят существенный вклад в значение коэффициента подобия. Коэффициент устанавливает регулярную эквивалентность между вершинами (см. п. 3.1).

Равенство (4.1) можно представить в виде

$$S_{Katz}(i,j) = \beta(C)_{i,j} + \beta^2(C^2)_{i,j} + \beta^3(C^3)_{i,j} + \cdots, \tag{4.2}$$

где C — матрица смежности графа. Таким образом, коэффициенты подобия для всех пар вершин можно представить в виде матрицы

$$S_{Katz} = (I - \beta C)^{-1} - I,$$

в которой элемент s_{ij} будет содержать коэффициент подобия вершин i и j (здесь I — единичная матрица). Метод дает хороший результат, если β достаточно большое, при этом $\beta < 1/\lambda$, где λ — наибольшее собственное значение матрицы C. Т. е. для обеспечения сходимости последовательности необходимо, чтобы матрица имела главный собственный вектор.

Следует отметить, что для неориентированных графов отношение подобия будет симметричным. Однако для орграфа отношение может не быть симметричным. Мы предполагаем, что в СЦС циклов не может быть, так как либо i цитирует j, либо верно обратное. Однако редкие исключения могут иметь место (см. Введение). Можно считать, что $S_{Katz}(i,j)$ отражает симметрию i по отношению к j. Необходимость симметрии в определении отношения подобия оспаривается в работе [35], где предложена теоретико-множественная модель подобия, учитывающая свойства акторов сети. Такой подход используется, например, в работе [36]. В любом случае, $S_{Katz}(i,j)$ может быть использован для ранжирования пар вершин, и в этом смысле он подобен вычислению дистанции, когда учитываются только кратчайшие пути от i до j и их длины, а в нашем случае учитываются все пути, их длины и количество. Вычисление коэффициента $S_{Katz}(i,j)$ требует больших затрат памяти и времени $\mathcal{O}(n^3)$, поэтому в подходящих случаях для больших сетей используются методы аппроксимации [15, 37].

Пример 4. Для пар вершин компоненты A_{16} по схеме (4.2) были вычислены коэффициенты $S_{Katz}(i,j)$ в режимах *in* и *out*. В обоих случаях сходимость достигается на шестом шаге, для режима $out\beta = 0{,}58$ (1/1,71), для режима $in\beta = 0{,}62$ (1/1,62). Результаты приведены в табл. 4.1.

Заключение. Для СЦС определены параметры, отражающие свойства пар вершин графа, и вычислены их значения. Результаты представлены в виде таблиц и рисунков. Следует заметить, что в литературе, например в работе [4], параметры пар вершин определены для неориентированных графов, поэтому для орграфов введены ограничения.

Таблица 3.5

Коэффициенты подобия пар компоненты A_{16} (первые 12 позиций)

Режим *in*							Режим *out*						
Ранг по *ККЦ*	Идентификаторы статей	*ККЦ* (3.2)	S_J (3.4)	S_A (3.5)	S_{AA} (3.6)	*Indeg* пары	Ранг по *КБС*	Идентификаторы по *КБС* статей	*КБС* (3.1)	S_J (3.4)	S_A (3.5)	S_{AA} (3.6)	*Outdeg* пары
1	78390, 117824	3	0,7500	0,2500	2,1455	4, 3	1–2	55774, 106436	3	0,4286	0,1429	2,2529	7, 3
2–6	78245, 821750	2	0,2857	0,1000	1,6316	4, 5	1–2	78243, 782440	3	0,7500	0,2500	1,9640	4, 3
2–6	78245, 117825	2	0,2857	0,1000	1,6316	4, 5	3–4	55774, 78242	2	0,2222	0,1714	1,6316	7, 4
2–6	78390, 117825	2	0,2857	0,1000	1,4241	4, 5	3–4	78242, 106436	2	0,4000	0,1667	1,6316	4, 3
2–6	82175, 117825	2	0,2500	0,0800	1,6316	5, 5	5–12	55774, 55775	1	0,1429	0,1429	1,4427	7, 1
2–6	117824, 117825	2	0,3333	0,1333	1,4241	3, 5	5–12	55774, 78243	1	0,1000	0,0357	0,6213	7, 4
7–12	55775, 58606	1	1,0000	1	0,5139	1, 1	5–12	55774, 78244	1	0,1111	0,0476	0,6213	7, 3
7–12	55775, 78241	1	1,0000	1	0,5139	1, 1	5–12	55774, 78389	1	0,1250	0,0714	0,7213	7, 2
7–12	55775, 78389	1	0,5000	0,5000	0,5139	1, 2	5–12	55774, 117824	1	0,1429	0,1429	0,6213	7, 1
7–12	55775, 78390	1	0,2500	0,2500	0,5139	1, 4	5–12	78241, 78242	1	0,2000	0,1250	0,7213	2, 4
7–12	55775,1 17824	1	0,3333	0,3333	0,5139	1, 3	5–12	78241, 78243	1	0,2000	0,1250	0,7213	2, 4
7–12	55775, 117825	1	0,2000	0,2000	0,5139	1, 5	5–12	78241, 78244	1	0,2500	0,1667	0,7213	2, 3

Таблица 4.1

Значения параметра $S_{Katz}\ (i,j)$ для компоненты A_{16} (первые 12 позиций)

Режим *in*				Режим *out*			
Ранг по S_{Katz}	Пары статей i,j	$S_{Katz}(i,j)$ (4.2)	Идентификаторы статей	Ранг по S_{Katz}	Пары статей i,j	$S_{Katz}(i,j)$ (4.2)	Идентификаторы статей
1	13,2	2,11	78243, 55775	1	1,14	1,70	55774, 82175
2	13,6	1,63	78243, 78241	2	11,14	1,40	78242, 82175
3	12,2	1,54	78244, 55775	3	1,5	1,31	55774, 78390
4	10,2	1,48	106436, 55775	4	1,16	1,29	55774, 78391
5	15,2	1,39	131949, 55775	5	1,7	1,22	55774, 117825
6	13,9	1,01	78243, 78245	6–12	1,4	0,92	55774, 78389
7–12	7,6	1,00	117825, 78241	6–12	4,16	0,92	78389, 78391
7–12	9,2	1,00	78245, 55775	6–12	5,14	0,92	78390, 82175
7–12	12,9	1,00	78244, 78245	6–12	10,7	0,92	106436, 117825
7–12	13,5	1,00	78243, 78390	6–12	11,9	0,92	78242, 78245
7–12	13,8	1,00	78243, 117824	6–12	12,14	0,92	78244, 82175
7–12	13,10	1,00	78243, 106436	6–12	13,14	0,92	78243, 82175

Главная компонента СЦС имеет диаметр, равный 34, и среднее расстояние между парами вершин, равное 9,6. Небольшое среднее расстояние между вершинами является одним из признаков "малого мира".

Рассмотренные параметры можно разделить на две группы. Для первой группы предполагается наличие пути из i в j (расстояние, размер (i-j)-разреза и коэффициент S_{Katz}); для второй предполагается наличие общих соседей (*КБС*, *ККЦ*, их нормированные варианты и коэффициент S_{AA}). Таким образом, множества пар, имеющих ненулевое значение параметров для этих групп, не совпадают. Ранжирование вершин существенно отличается в зависимости от используемого параметра, что указывает на то, что преимущество использования конкретного параметра зависит от задачи. Так, для предсказания будущих связей между вершинами предпочтительнее подобие по Кацу. Для второй группы вычислены коэффициенты ранговой корреляции, из значений которых видно, что наиболее близкими оказались *КБС*/*ККЦ* и коэффициент S_{AA}, ненормированные коэффициенты подобия более других зависят от степени вершины.

Подробно рассмотрена компонента из шестнадцати вершин A_{16}. Выяснилось, что ранги пар согласно размеру (i-j)-разреза и коэффициенту S_{Katz} практически совпадают. Исследование показало, что среди путей, ведущих из i в j, почти все ребро-независимы, что явилось причиной совпадения [38].

Параметры первой группы отражают распространение информации путем цитирования, а параметры второй группы — семантическую близость публикаций. Они могут служить основой для выявления сообществ внутри сети, что является предметом дальнейшего исследования.

Список литературы

1. Бредихин С. В., Ляпунов В. М., Щербакова Н. Г., Юргенсон А. Н. Параметры "центральности" узлов сети цитирования научных статей // Проблемы информатики. 2016. № 1. С. 39–57.

2. RePEc. General principles. [Electron. resource]. http://repec.org/.

3. Харари Ф. Теория графов. М.: Мир. 1973.

4. Оре О. Теория графов. М.: Наука. 1980. 338 с.

5. Кормен Т., Лейзерсон Ч., Ривест Т. Алгоритмы: построение и анализ. М.: МЦНМО. 2002. 960 с.

6. Csárdi G., Nepusz T. The igraph software package for complex network research // InterJournal Complex Systems. 2006. 1695 P. [Electron. resource]. http://igraph.org/r/doc/.

7. Batagelj V., Mrvar A. Pajek — Program for large network analysis [Electron. resource]. http://vlado.fmf.uni-lj.si/pub/networks/doc/pajek.pdf.

8. Network analysis (Ed. Brandes U., Erlebach T.) Berlin: Springer-Verlag. 2005. 471 p.

9. Provan J. S., Shier D. R. A paradigm for listing (s,t)-cuts in graphs // Algorithmica. 1996. V. 15. P. 351–372.

10. Ford L. R., Fulkerson D. R. Maximal flow through network // Canad. J. Math. 1956. V. 8. P. 399–404.

11. Goldberg A. V., Tarjan R. E. A new approach to the maximum-flow problem // J. of the ACM. 1988. V. 35, iss. 4. P. 921–940.

12. Kannan R., Vempala S., Vetta A. On clusterings — good, bad and spectral // Proc. of the 41st Annual IEEE Symposium on Foundations of Computer Science (FOCS'00). 2000. P. 367–378. [Electron. resource]. http://www.cc.gatech.edu/∼vempala/papers/jacm-spectral.pdf.

13. Flake G. W., Tarjan R. E., Tsioutsiouliklis K. Graph clustering and minimum cut trees // [Electron. resource]. http://www.cs.princeton.edu/∼kt/imath.pdf.

14. Görke R., Hartmann T., Wagner D. Dynamic graph clustering using minimum-cut trees // J. of Graph Algorithms and Applications. 2012. V. 16, N 2. P. 411–446.

15. The link prediction problem for social networks // Proceedings of the 12th International Conference on Information and Knowledge Management, 2003. (Сокращ. версия). [Electron. resource]. http://www.cs.carleton.edu/faculty/dlibenno/papers/link-prediction/link.pdf.

16. Huang Z., Lin D. K. J. The time-series link prediction problem with applications in communication surveillance // INFORMS J. Computing. 2009. V. 21. P. 286–303.

17. Garfield E. Citation Indexing: A natural science literature retrieval system for the social sciences // The American Behavioral Scientist. 1964. V. 7, iss.10. P. 58–61.

18. Garner R. A computer-oriented graph theoretic analysis of citation index structure // Three Drexel information science studies. Philadelphia: Drexel Institute of Technology. 1965. P. 1–46.

19. Jarneving B. A comparison of two bibliometric methods for the mapping of the research front // Scientometrics. 2005. V. 65, iss. 2. P. 245–263.

20. Glänzel W., Czerwon H. J. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level // Scientometrics. 1996, V. 32, iss. 2. P. 195–221.

21. Garfield E. Mapping the structure of science. Citation Indexing: Its Theory and Applications in Science, Technology, and Humanities. NY: Wiley. 1979. P. 98–147. [Electron. resource]. http://www.garfield.library.upenn.edu/ci/chapter8.pdf.

22. Börner K., Chen C., Boyack K. W. Visualizing knowledge domains // Annual Rev. Inform. Sci. Technol. (ARIST).

23. Boyack K. W., Klavans R., Börner K. Mapping the backbone of science // Scientometrics. 2005. V. 64, iss. 3. P. 351–374.

24. Newman M. E. J. Clustering and preferential attachment in growing networks // Physical Review E. 2001. V. 64, iss. 2. 025102.

25. Newman M. E. J. Networks. An introduction. NY: Oxford University Press. 2010. 772 P.

26. Kessler M. M. Bibliographic coupling between scientific papers // Amer. Documentation. 1963. V.14, iss.1. P. 10–25.

27. Маршакова И. В. Система связей между документами, построенная на основе ссылок: по данным Science Citation Index // НТИ. Сер.: 2. 1973. № 6. С 3–8.

28. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents // J. Amer. Soc. Inform. Sci 1973. V. 24, iss. 4. P. 265–269.

29. Salton G. Associative document retrieval techniques using bibliographic information // J. of the ACM. 1963. V. 10, iss. 4. P. 440-457.

30. Van Eck N. J.,Waltman L., Van den Berg J., Kaymak U. Visualizing the computational intelligence field // IEEE Computational Intelligence Magazine. 2006. V. 1, iss. 4. P. 6–10.

31. Rip A., Courtial J.-P. Co-word maps of biotechnology: An example of cognitive scientometrics // Scientometrics. 1984. V. 6, iss. 6. P. 381–400.

32. Van Eck N. J., Waltman L. How to normalize cooccurrence data? An analysis of some well-known similarity measures // J. Amer. Soc. Inform. Sci. Tech. 2009. V. 60, iss. 8. P. 1635–1651.

33. Adamic L. A., Adar E. Friends and neighbors on the web // Soc. Networks. 2003. V. 25, iss. 3. P. 211–230.

34. Katz L. A new status index derived from sociometric analysis // Psychometrika. 1953. V. 18. P. 39–43.

35. Tversky A. Features of similarity // Psychological Review. 1977. V. 84. P. 327–352.

36. Chen H., Giles C. L. ASCOS: An asymmetric network structure context similarity measure // Proc. of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara, ON, Canada, August 25–29. P. 442–449.

37. Acar E., Dunlavy D., Kolda T. Link prediction on evolving data using matrix and tensor factorizations // IEEE International Conference on Data Mining Workshops, 2009. Miami, Florida, December 6. P. 262–269.

38. Menger K. Zur allgemeinen Kurventheorie // Fund. Math. 1927. V. 10. P. 96–115.

Бредихин Сергей Всеволодович — канд. техн. наук, зав. лабораторией Института вычислительной математики и математической геофизики СО РАН;
e-mail: bred@nsc.ru;
Ляпунов Виктор Михайлович — ведущий инженер Института вычислительной математики и математической геофизики СО РАН;
e-mail: vic@nsc.ru;
Щербакова Наталья Григорьевна — ст. науч. сотр. Института вычислительной математики и математической геофизики СО РАН;
e-mail: nata@nsc.ru.

Дата поступления — 22.03.2016