

МЕРА „ЦЕНТРАЛЬНОСТИ“ ДЛЯ РАНЖИРОВАНИЯ 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

Предметом изучения является сеть цитирования, отражающая взаимосвязь журнальных статей из коллекции базы данных *RePEc*. Изучены характеристики множества публикаций и вычислена мера „центральность по посредничеству“ вершин взвешенного графа коцитирования для нормированного и ненормированного случаев. Выполнено ранжирование коллекции статей на основе этой меры.

Ключевые слова: сети цитирования и коцитирования, мера „центральность по посредничеству“, алгоритм Брандеса.

Subject of studying is the citation network reflecting the structure of interrelations between journal articles of the database *RePEc* collection. Characteristics of a publication set are studied and betweenness centralities for vertices of the weighted cocitation graph are found for normalized and unnormalized cases. The journal articles ranking based on that measure is executed.

Key words: citation and cocitation networks, betweenness centrality measure, Brandes algorithm.

1. Определения. „Центральность актора“ является ключевой концепцией сетевого анализа, предназначенной для определения значимости позиции актора в топологии сети. Интуитивные подходы к определению центральности применительно к социальным сетям исследованы в работе [1]. Ряд важных мер центральности, рассмотренных с точки зрения прохождения потоков трафика через сеть, представлен в работе [2]. А именно, траекторий прохождения (кратчайшие пути, цепи или маршруты) и способа передачи информации (широкое вещание, копирование или трансляция по выбранному маршруту). Так, рассматриваемая здесь „центральность по посредничеству“ (betweenness centrality, далее C_B) основана на идее передачи сообщений между вершинами информационной сети по кратчайшему пути, выбранному случайным образом [3].

Как показано в работе [4], на основании данных о цитировании документов можно построить сеть цитирования и представить ее в виде ориентированного графа. Рассмотрим множество D , состоящее из n документов, на котором задано отношение $R \subseteq D \times D$

$$d_1 R d_2 \equiv d_1 \text{ цитирует } d_2,$$

которое определяет сеть цитирования $N(D, R)$. Здесь под термином „цитирует“ будем понимать наличие цитируемого документа в списке ссылок цитирующего. В графе $G = (V, E)$, представляющем сеть цитирования, акторы (вершины) соответствуют документам, а множество E состоит из направленных ребер (дуг), таких что дуга $(d_1, d_2) \in E$ тогда и только тогда, когда выполняется отношение $d_1 R d_2$. Этот ориентированный ациклический граф без кратных ребер называется „графом цитирования“ или „сетью цитирования“.

Матрица смежности графа G , квадратная матрица C размерности n , такая, что элемент этой матрицы $c_{ij} = 1$, если документ d_i процитировал документ d_j , названа „матрицей цитирования“.

На основе данного графа G может быть построено несколько графов путем наложения ограничений на множество D (например, только документы, относящиеся к определенному временному периоду) и на отношение R (например, цитирования только между документами определенного типа). Преобразование графа путем объединения нескольких вершин в одну (например, по принципу принадлежности к одному периодическому изданию) дает сеть цитирования новых акторов. На множестве D можно также определить новое отношение.

Построим отношение R_1 , такое что

$$d_1 R_1 d_2 \equiv (\exists d_3) d_3 R d_1 \& d_3 R d_2.$$

Это отношение названо отношением „коцитирования“ (*co-citation, cocitation*) [5, 6]. Два документа находятся в отношении коцитирования, если существует хотя бы один документ, цитирующий оба эти документа. Частота появления пары документов (d_1, d_2) в списках литературы различных документов характеризует степень взаимосвязи между ними. Согласно [7], будем называть эту частоту коэффициентом коцитирования (ККЦ). Таким образом, ККЦ документов d_1 и d_2 — это мощность пересечения множеств $\Omega_1 \cap \Omega_2$, где Ω_1 и Ω_2 — множества документов, цитирующих документы d_1 и d_2 соответственно. Отношение R_1 может рассматриваться как отношение близости между парой документов.

Отношение R_1 определяет сеть коцитирования (СКЦ) $N(D, R_1)$. В представлении сети в виде неориентированного графа отношению $d_1 R_1 d_2$ соответствует ребро (d_1, d_2) . Если сила связи между документами не учитывается, граф будет невзвешенным, в противном случае в качестве веса может выступать ККЦ ($w = \text{ККЦ}$). Формула вычисления ККЦ для документов d_i и d_j , ($i \neq j$):

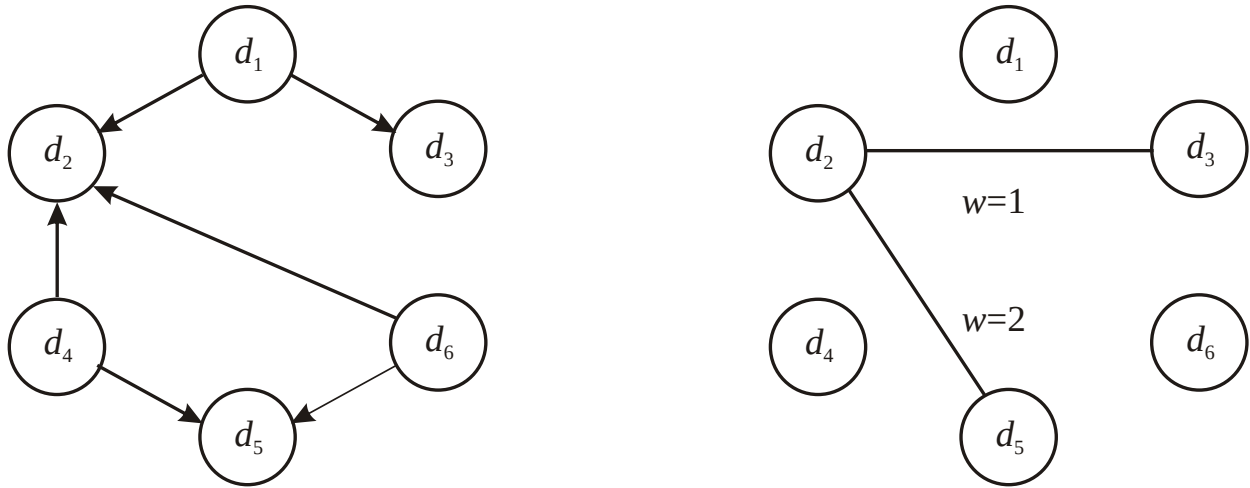
$$\sum_{k=1}^n c_{ki} c_{kj} = (C^T C)_{ij}.$$

Матрица смежности взвешенного графа коцитирования CC может быть получена из матрицы смежности C графа цитирования:

$$CC' = C^T C$$

с последующим обнулением главной диагонали матрицы CC' (так как $(CC')_{ii}$ — это количество цитирований документа d_i). На рис. 1 приведен пример сети цитирования и сети коцитирования, построенных на одних и тех же данных о цитировании: документ d_1 цитирует документы d_2 и d_3 , документ d_4 цитирует документы d_2 и d_5 , документ d_6 также цитирует документы d_2 и d_5 .

2. Центральность по посредничеству. Пусть $G = (V, E)$ — связный неориентированный граф и w — функция взвешивания ребер, такая что $w(e) > 0$ ($e \in E$) для взвешенных графов и $w(e) = 1$ для невзвешенных графов. Определим путь от вершины $s \in V$ до вершины $t \in V$ как последовательность несовпадающих вершин и ребер $(v_1, e_1, v_2, e_2, \dots, e_{n-1}, v_n)$, начинающуюся в s ($s = v_1$) и заканчивающуюся в t ($t = v_n$), такую что каждое ребро соединяет предшествующую вершину с последующей. Если в графе нет кратных ребер, то путь можно представлять в виде последовательности вершин. Длина пути p — это сумма весов его ребер:

Рис. 1. a — сеть цитирования; b — сеть коцитирования

$$w(p) = \sum_{i=1}^{n-1} w(e_i).$$

Расстояние между вершинами s и t , $d_G(s, t)$ — это минимум среди длин путей, соединяющих s и t . По определению $d_G(s, s) = 0$, $d_G(s, t) = d_G(t, s)$. Путь между вершинами s и t , длина которого равна расстоянию между s и t , называется кратчайшим.

Пусть σ_{st} — количество кратчайших путей от вершины $s \in V$ до вершины $t \in V$. Поскольку граф неориентированный, $\sigma_{st} = \sigma_{ts}$. Для неориентированного случая через σ_{st} будем обозначать количество путей между s и t $((\sigma_{st} + \sigma_{ts})/2)$. Пусть $\sigma_{st}(v)$ — количество кратчайших путей между s и t , проходящих через $v \in V$. Тогда индекс $C_B(v)$ [3] для вершины v определяется следующим образом:

$$C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}.$$

Обозначим через

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$$

долю кратчайших путей между s и t , проходящих через v . Эту величину, которую назовем парной зависимостью (s и t от v), можно также интерпретировать как вероятность того, что случайно выбранная связь между s и t проходит через v . Для адаптации определения в случае, когда не существует пути между s и t (числитель $\sigma_{st}(v)$ и знаменатель σ_{st} равны нулю), принято соглашение: $\delta_{st}(v) = 0$.

3. Метод. С использованием определения парной зависимости меру „центральности“ вершины v можно представить следующим образом:

$$C_B(v) = \sum_{s \neq t \neq v \in V} \delta_{st}(v).$$

Процедура вычисления традиционно состоит из двух шагов: а) найти все кратчайшие пути между всеми парами вершин; б) просуммировать все парные зависимости.

В работе [8] предложен подход, в котором техника аккумуляирования значений центральности интегрирована с поиском кратчайших путей. Подход основан на рекурсивной формуле, устанавливающей связь между зависимостями вершин и их предшественников.

Зависимость (*dependency*) единичной вершины $s \in V$ от вершины $v \in V$ определяется как

$$\delta_s(v) = \sum_{t \in V} \delta_{st}(v). \quad (1)$$

Тогда $C_b(v)$ можно представить в виде:

$$C_B(v) = \sum_{s \in V} \delta_s(v). \quad (2)$$

Обозначим через $P_s(v)$ множество предшественников (*predecessors*) вершины $v \in V$ на кратчайших путях из $s \in V$:

$$P_s(v) = \{u \in V : (u, v) \in E, d_G(s, v) = d_G(s, u) + w(u, v)\}.$$

Связь количества кратчайших путей от s до v ($s \neq v \in V$) с количеством путей до предшественников:

$$\sigma_{sv} = \sum_{u \in P_s(v)} \sigma_{su}. \quad (3)$$

Для вычисления значения зависимости вершины $s \in V$ от любой вершины $v \in V$ можно применять следующую ключевую рекурсивную формулу:

$$\delta_s(v) = \sum_{w : v \in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}} (1 + \delta_s(w)). \quad (4)$$

Алгоритм вычисления центральности всех вершин основывается на вычислении для любой вершины $s \in V$ зависимостей от остальных вершин. Для этого выполняются два шага:

а) производится подсчет длин и количества кратчайших путей от $s \in V$ до всех остальных вершин и запоминается состав путей;

б) вычисляются зависимости s от всех вершин на путях с использованием (1), (3), (4) и суммируются со значением центральности, согласно (2).

В результате итерации будут рассмотрены все пары вершин и просуммированы парные зависимости. Граф неориентированный, поэтому значение делится на два, так как кратчайшие пути между парой вершин рассматриваются дважды. С деталями можно ознакомиться в работе [9], там же приведен алгоритм вычисления C_B для взвешенных графов.

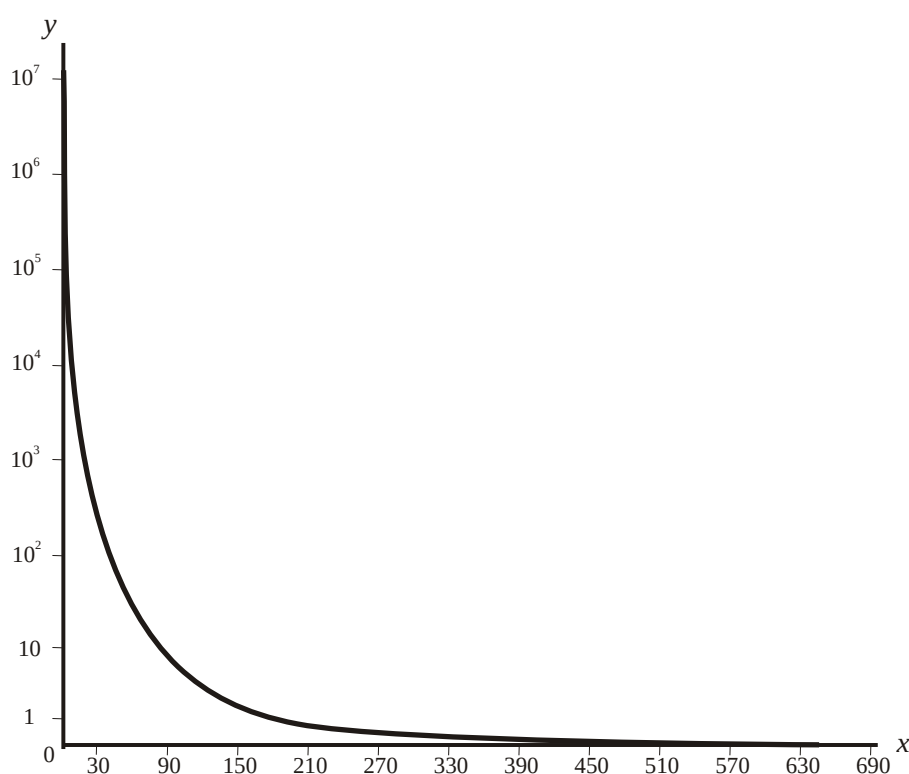


Рис. 2. Зависимость между значением ККЦ и количеством пар, коцитируемых с данной частотой

4. Задача и данные. Задачей настоящей работы является вычисление значения меры C_B для вершин графа, представляющего СКЦ, построенную на информации о цитировании, извлеченной из распределенной базы библиографических данных (далее РББД) *RePEc* [10]. Целью является ранжирование (установление отношения линейного порядка) множества статей из научных журналов на основе этой меры. В РББД на момент извлечения насчитывалось 5811 коллекций документов по экономике, из них 1708 журналов. Общее число статей в журналах — 435 026; из них состоящих в отношении коцитирования — 323 159. Число пар статей, состоящих в отношении коцитирования — 13 875 461, максимальный ККЦ = 687 имеет одна пара статей: Arellano M., Bond S. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations // *Review of Economic Studies*. 1991. Blundell R., Bond S. Initial conditions and moment restrictions in dynamic panel data models // *Journal of Econometrics*. 1998.

На рис. 2 представлена зависимость между значением ККЦ и количеством пар статей, коцитируемых с данной частотой. По оси x — значение ККЦ, по оси y — количество пар статей в логарифмическом масштабе. Для аппроксимации точечных значений в предположении, что зависимость имеет степенной характер ($f(x) = ax^{-\lambda}$), использовался метод наименьших квадратов. В данном случае $\lambda = 3,171$.

5. Вычислительный эксперимент. Для сокращения времени вычислительного эксперимента рассмотрим несколько вариантов редуцирования данных.

5.1. Порог по цитированию. В начале эксперимента был введен порог на цитируемость, равный 58. В результате количество рассматриваемых статей сократилось до 4008. Считаем это множество исходным и нем рассматриваем отношения цитирования и коци-

Таблица 1

Статьи множества P_1 , получившие наивысший ранг

Ранг	C_B	Статья
1	595 563	La Porta R., Lopez-de-Silanes F., Shleifer A., Vishny R. W. Law and Finance // J. of Political Economy. 1998
2	472 148	Newey W. K., West K. D. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix // Econometrica. 1987
3	334 876	Hansen L. P. Large Sample Properties of Generalized Method of Moments Estimators // Econometrica. 1982
4	285 395	Bollerslev T. Generalized autoregressive conditional heteroskedasticity // J. of Econometrics. 1986
5	282 655	White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity // Econometrica. 1980
6	214 823	Arellano M., Bond S. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations // Review of Economic Studies. 1991
7	213 099	Levine R., Zervos S. Stock Markets, Banks, and Economic Growth // American Economic Review. 1998
8	199 276	Jensen M. C. Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers // American Economic Review. 1986
9	172 694	Fama E. F., MacBeth J. D. Risk, Return, and Equilibrium: Empirical Tests // J. of Political Economy. 1973
10	165 942	Fama E. F., French K. R. Common risk factors in the returns on stocks and bonds // J. of Financial Economics. 1993
11	163 688	Lucas R. Jr. On the mechanics of economic development // J. of Monetary Economics. 1988
12	148 304	Mehra R., Prescott E. C. The equity premium: A puzzle // J. of Monetary Economics. 1985

тирования. В рамках этого множества 2512 статей состоят в отношении коцитирования. Количество коцитируемых пар — 45 214, максимальный ККЦ = 17 имеет одна пара статей: Levin R., Zervos S. Stock markets, banks, and economic growth // American Economic Rev. 1998. La Porta R., Lopez-de-Silanes F., Shleifer A., Vishny R. W. Law and Finance // J. of Political Economy. 1998.

Рассматривается ненаправленный взвешенный граф G с 2512 вершинами. Этот граф распадается на несколько связанных компонент, первая из которых содержит 2479 вершин, остальные — не более пяти вершин. Обозначим первую компоненту через G_1 , а множество статей, соответствующих вершинам графа, через P_1 . Вычислим значения центральности по посредничеству для вершин этого графа. Нас интересуют пути с большой суммой коцитирования, а при вычислении центральности по посредничеству рассматриваются кратчайшие пути, поэтому в качестве веса ребра используется значение $1/\text{ККЦ}$. Статьи с наивысшим значением C_B приведены в табл. 1.

5.2. *Порог по коцитированию.* Введем порог по коцитированию для всего множества журнальных статей, равный 30. Количество статей, для которых ККЦ $\geq$

30, равно 2976, количество пар — 6533. Введение порога на цитирование привело к тому, что связный граф распался на ряд компонент. В работе [11] эти компоненты интерпретируются как научные специальности или разделы специальностей. В нашем случае получилось множество небольших компонент и одна компонента, состоящая из 1981 вершины. Обозначим эту компоненту через G_2 , а множество статей, соответствующих вершинам графа, через P_2 . Вычислим значения C_B для вершин графа G_2 . Статьи с наивысшим значением меры приведены в табл. 2.

Таблица 2

Статьи множества P_2 , получившие наивысший ранг

Ранг	C_B	Статья
1	535 937	Fama E. F., French K. R. Common risk factors in the returns on stocks and bonds // J. of Financial Economics. 1993
2	454 085	Jensen M. C., Meckling W. H. Theory of the firm: Managerial behavior, agency costs and ownership structure // J. of Financial Economics/ 1976
3	442 380	Arellano M., Bond S. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations // Review of Economic Studies. 1991
4	388 977	La Porta R., Lopez-de-Silanes F., Shleifer A., Vishny R. W. Law and Finance // J. of Political Economy. 1998
5	345 607	Newey W. K., West K. D. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix // Econometrica. 1987
6	322 555	Bollerslev T. Generalized autoregressive conditional heteroskedasticity // J. of Econometrics. 1986
7	311 038	Mankiw N. G., Romer D., Weil David. N. A Contribution to the Empirics of Economic Growth // The Quarterly J. of Economics. 1992
8	277 215	Johansen S. Statistical analysis of cointegration vectors // J. of Economic Dynamics and Control. 1988
9	271 734	Hansen L. P. Large Sample Properties of Generalized Method of Moments Estimators // Econometrica. 1982
10	242 100	Black F., Scholes, M. S. The Pricing of Options and Corporate Liabilities // J. of Political Economy. 1973
11	193 761	Im K. S. Pesaran M. H., Shin Y. Testing for unit roots in heterogeneous panels // J. of Econometrics; Elsevier; 2003
12	185 934	Kahneman D., Tversky A. Prospect Theory: An Analysis of Decision under Risk // Econometrica. 1979

При сравнение таблиц 1 и 2 видно, что 7 из 12 статей присутствуют в обеих таблицах. Если рассмотреть 20 статей с наибольшим значением C_B , то 10 из 20 статей попадают в обе таблицы.

5.3. *Нормирование при вычислении ККЦ.* На значение ККЦ влияет количество цитирований, так как вероятность иметь большее значение коэффициента выше у пары с большим количеством цитирований. При рассмотрении отношения подобия объектов принято нормировать сырые данные [12] и [13]. Существует ряд приемов нормирования, от

Таблица 3

Статьи, получившие наивысший ранг в нормированном случае

Ранг	C_B	Статья
1	425 380	Loughran T., Ritter J. R. The New Issues Puzzle // J. of Finance. 1995
2	395 680	Gompers P., Ishii J., Metrick A. Corporate Governance And Equity Prices // The Quarterly J. of Economics, 2003
3	391 108	Geert B., Campbell R. H. Foreign Speculators and Emerging Equity Markets // J. of Finance. 2000
4	381 992	Rajan R. G., Zingales L. Financial Dependence and Growth // American Economic Review. 1998
5	380 188	La Porta R., Lopez-de-Silanes F., Shleifer A., Vishny R. W. Law and Finance // J. of Political Economy. 1998
6	378 300	Merton R. C. An Intertemporal Capital Asset Pricing Model // Econometrica. 1973
7	378 218	Mehra R., Prescott E. C. The equity premium: A puzzle // J. of Monetary Economics. 1985
8	370 688	Bekaert G., Harvey C. R., Lundblad Ch. Does financial liberalization spur growth? // J. of Financial Economics. 2005
9	322 224	Calvo G. A. Staggered prices in a utility-maximizing framework // J. of Monetary Economics. 1983
10	304 130	Breeden D. T. An intertemporal asset pricing model with stochastic consumption and investment opportunities // J. of Financial Economics. 1979
11	303 456	Hayne E. L. Agency Costs, Risk Management, and Capital Structure // J. of Finance. 1998
12	299 186	Ferson W. E., Harvey C. R. The Risk and Predictability of International Equity Returns // Review of Financial Studies. 1993

выбора которых зависит результат исследования. В работе [14] проведен анализ четырех часто используемых методов нормирования:

а) коэффициент ассоциативности

$$S_A(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i s_j}, \quad (5)$$

б) „косинусная“ мера

$$S_C(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\sqrt{s_i s_j}},$$

в) индекс включаемости

$$S_I(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\min(s_i, s_j)},$$

г) индекс Жаккара

$$S_J(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i + s_j - c_{ij}},$$

где s_i — встречаемость объекта i (в данном — случае количество цитирований), c_{ij} — совместная встречаемость объектов i и j (в данном случае — коэффициент коцитирования). Показано, что имеется существенное различие между коэффициентом ассоциативности и остальными мерами. Коэффициент ассоциативности лучше корректирует эффект размера. Поэтому для нормирования наших данных используется метод (5).

При сравнение таблиц 2 и 3 видно, что только одна из 12 статей присутствует в обеих таблицах. Из 20 статей с наибольшим значением C_B две статьи попадают в обе таблицы. Пересечение трех таблиц — это статья La Porta R., Lopez-de-Silanes F., Shleifer A., Vishny R. W. Law and Finance // J. of Political Economy. 1998, имеющая ранги 1, 4 и 5 соответственно. Для сравнения значений C_B статей коллекции P_2 в нормированном и ненормированном случаях вычислен коэффициент корреляции Пирсона (0,315), а также коэффициент ранговой корреляции Спирмена (0,868).

Заключение. В работе приведен опыт ранжирования журнальных статей на основе данных об их цитировании с учетом структурных особенностей сети коцитирования, а именно, „центральности“ ее вершин. Используется мера „центральность по посредничеству“, для вычисления которой применен метод Брандеса. Показано влияние способов фильтрации данных на результаты ранжирования.

Список литературы

1. FREEMAN L. C. Centrality in social networks. Conceptual clarification // Social networks. 1978/79. V. 1. P. 215–239.
2. BORGATTI S. P. Centrality and Network Flow // Social Networks. 2005. V. 27. P. 55–71.
3. FREEMAN L. C. A set of measures of centrality based upon betweenness // Sociometry. 1977. V. 40. P. 35–41.
4. PRICE D. A general theory of bibliometric and other cumulative advantage process // J. Amer. Soc. Inform. Sci. 1976, N 27. P. 292–306.
5. МАРШАКОВА И. В. Система связей между документами, построенная на основе ссылок: по данным Science Citation Index // НТИ. Сер. 2. 1973. № 6. С. 3–8.
6. 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.
7. БРЕДИХИН С. В., КУЗНЕЦОВ А. Ю. Методы библиометрии и рынок электронной научной периодики. Новосибирск, Москва: ИВМиМГ СО РАН, НЭЙКОН, 2012. 254 с.
8. BRANDES U. A faster algorithm for betweenness centrality // J. of Mathematical Sociology. 2001. V. 25, iss. 2. P. 163–177.
9. БРЕДИХИН С. В., ЛЯПУНОВ В. М., ЩЕРБАКОВА Н. Г. Мера важности научной периодики — „центральность по посредничеству“ // Пробл. информ. 2014. № 3. С. 53–63.
10. REPEC. General principles. [Electron. Resource]. <http://repec.org/>.
11. SMALL H., GRIFFITH B. C. The structure of scientific literatures. 1: Identifying and graphing specialties // Sci. Studies. 1974. V. 4. P. 17–40.
12. LEYDESDORFF L. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index // J. Amer. Soc. Inform. Sci. Tech. 2008. V. 59, iss. 1. P. 77–85.
13. UPHAM S. P., SMALL H. Emerging research fronts in science and technology: patterns of new knowledge development // Scientometrics. 2010. V. 83, iss. 1. P. 15–38.

14. 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.

*Бредихин Сергей Всеволодович — канд. техн. наук,
зав. лабораторией Ин-та вычислительной математики
и математической геофизики СО РАН; e-mail: bred@nsc.ru;*

*Ляпунов Виктор Михайлович — ведущий инженер
Ин-та вычислительной математики и математической
геофизики СО РАН; e-mail: vic@nsc.ru;*

*Щербакова Наталья Григорьевна — ст. научн. сотр.
Ин-та вычислительной математики и математической
геофизики СО РАН; e-mail: nata@nsc.ru.*

Дата поступления — 30.01.2015