ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 27 МАТЕМАТИКИ Выпуск 2
2020

А. С. Козицын (Москва, МГУ им. М. В. Ломоносова)**. Способ расширения тематического поиска журналов.**

УДК 004.912

Резюме: В работе рассматривается метод увеличения полноты поиска журналов при проведении тематического анализа на основе графа соавторства.

Ключевые слова: граф соавторства, тематический анализ, библиография, наукометрия.

В настоящее время в мире издается огромное количество научных журналов. Например, в Web of Science зарегистрировано более 22 тысяч журналов, в Scopus — более 23 тысяч, система РИНЦ содержит данные более чем о 68 тысячах научных журналов. Это делает процесс ручного поиска и тематического сопоставления и журналов очень трудоемким . Еще более сложная ситуация в области тематического сопоставления конференций. В отличии от журналов, для конференций не существует единого каталога, который бы охватывал значительную часть всех проводимых в мире конференций. Необходимо развитие аналитический инструментов, позволяющих определять тематическую близость журналов и конференций в автоматическом режиме. Такие механизмы могут использоваться для автоматизированного создание онтологий[1], [2], в том числе в целях построения правил разграничения доступа [3], а также для предоставления рекомендаций научным сотрудникам о возможных новых местах представления своих результатов.

Одним из способов определения тематической близости является анализ графа соавторства. В работе [4] описан алгоритм определения тематической близости журналов на основе наличия в разных журналах публикаций одного автора. Результатом работы алгоритма является матрица расстояний между журналами, которая используется в настоящий момент для поиска похожих журналов в системе ИСТИНА.

В настоящей работе представлен метод расширения результатов поиска за счет построения и оценки стоимости путей в графе соавторства, содержащих более чем две вершины.

Формальная постановка задачи. Дано множество авторов A, множество журналов J и функция $F : A \times J \to \mathbb{R}$, описывающая коэффициент публикационной активности автора в журнале. Требуется построить функцию $T : J \times J \to \mathbb{R}$, отражающую степень тематической близости журналов.

На первом шаге алгоритма, как и в работе [4] производится построение матрицы близости W, элементы которой вычисляются на основе функции F. Как показано в работе [4], эту матрицу можно использовать для определения тематической близости журналов. На настоящий момент именно она используется для определения близких по тематике журналов в системе ИСТИНА. Однако, полнота такого решения оказывается недостаточна. Эксперименты на реальных данных показали, что учет путей большей длины в графе связей журналов, заданного множеством вершин J и множеством ребер $(j_1, j_2) \in J \times J$, $T(j_1, j_2) > 0$, позволяет увеличить полноту результатов поиска, и не сильно снижая его точность.

Следует отметить, что использование непосредственно функции T для вычисления весов путей дает плохие результаты из-за наличия междисциплинарных журналов, которые создают ошибочные связи между журналами различной тематики. С целью уменьшения влияния таких журналов перед проведением вычисления путей производится нормировка расстояний и построение новой матрицы близости $\widehat{T}$. В рамках настоящих исследований рассматривалось два способа нормировки: по сумме $\widehat{t}_{ij} = \frac{t_{ij}}{\sum_i t_{ij}}$ и по количеству $\widehat{t}_{ij} = \frac{t_{ij}}{\sum_i \delta(t_{ij})}$, где

$t_{ij}^{(0)} = 1$, при $i = j$
$t_{ij}^{(0)} = \widehat{t}_{ij}$, при $i \neq j$
$t_{ij}^{(n)} = \sum_k \widehat{t}_{ik} t_{kj}^{(n-1)}$

Тестирование программной реализации алгоритма проводилось на данных системы ИСТИНА. Были отобраны работы, входящие в top80 новых работ, и не имеющие положительных связей в исходной матрице T, при $n \in 1, 2, 3$ с критерием по сумме или по количеству. Сравнение результатов проводилась при помощи подсчета суммарного коэффициента $s_i = k_i(80 - r_i)$, где k_i — экспертная оценка степени правильности найденной тематической связи по шкале $0 - 2$, а r_i — позиция связи в отсортированном списке top80.

Номер шага	1	2	3
Нормировка по сумме	2528	2631	2589
Нормировка по количеству	2458	2540	2226

Таким образом, наилучший результат получился для коэффициента с нормировкой по сумме на втором шаге расширения.

Работа выполнена при финансовой поддержке РФФИ (грант № 18-07-01055)

СПИСОК ЛИТЕРАТУРЫ

1. *Афонин С. А.*, *Козицын А. С.*, *Шачнев Д. А.* Программные механизмы агрегации данных, основанные на онтологическом представлении структуры реляционной базы наукометрических данных. — Программная инженерия, 2016, т. 7, № 9, с. 408–413.
2. *Платонов А. В.*, *Полещук Е. А.* Методы автоматического построения онтологий. — Программные продукты и системы, 2016, № 2, с. 47–52.
3. *Afonin S.* Ontology models for access control systems. — Proc. of the 3rd International Conference Russian-Pacific Conference on Computer Technology and Applications (RPC), 2018, p. 1–6.
4. *Козицын А. С.*, *Афонин С. А.*, *Шачнев Д. А.* Определение тематической близости научных журналов и конференций с использованием анализа графа соавторства. — Электронные библиотеки, т. 23, № 3, с. 514–525.

UDC 004.912

Kozitsyn A. S. (Moscow, Russia, Lomonosov Moscow State University)**. Method for expanding the thematic journals search area**

Abstract: The paper considers a method for increasing the completeness of the search for journals when conducting thematic analysis based on the co-authorship graph.

Keywords: thematic analysis, scientometrics, bibliography, co-authorship graph.