ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 24 МАТЕМАТИКИ Выпуск 5
2017

А. С. Козицын (Москва, НИИ механики МГУ)**. Алгоритм определения авторов статей по библиографическим данным.**

Управление большими организациями науки и образования невозможно без внедрения систем автоматизации сбора данных о научной деятельности сотрудников. Одним из важнейших показателей эффективности научной деятельности сотрудников организации является количество публикаций. В рамках автоматизации сбора библиографических данных о публикациях необходимо решать задачу определения автора по указанным в публикации фамилии и инициалам. Для распространенных фамилий контекстного поиска недостаточно и необходимо проводить более сложный анализ.

Один из методов решения этой задачи, основанный на поиске максимально связанных подграфов в графе соавторства, описан в работах [1]. Метод достаточно эффективен, однако, он имеет достаточно большую вычислительную сложность, и не использует информацию об авторизации пользователя.

Разработанный автором этой работы алгоритм на первом шаге выделяет список возможных авторов для каждой фамилии, упоминающейся в библиографическом описании статьи, с учетом всех вариантов написания его фамилии и инициалов, встречавшихся ранее. Такой анализ необходим для работы со статьями, изданными на других языках. Если среди возможных авторов встречается авторизованный в настоящий момент пользователь, то он считается определенным. Далее производится сортировка по количеству вариантов для каждой фамилии и, начиная с наименее частотных фамилий, для пар фамилий осуществляется оценка вероятности соавторства для каждой пары авторов, соответствующих этим фамилиям. Для каждого последующего ФИО выбирается автор с наилучшим ребром связи с предыдущими. Если остаются нераспознанные, то определяется лучшее ребро для каждой пары вариантов.

Тестирование алгоритма проводилось на графе соавторства, имеющего около 226 тысяч вершин (авторов) и 5 миллионов ребер. Для построения графа соавторства использовалась информация из статей, тезисов, книг и проектов по данным системы ИСТИНА[2]. В ходе тестирования, было обработано 540 тысяч статей. Совпадение результатов расчета алгоритма с реальными данными составило 520 тысяч записей. Дальнейшее улучшение результатов возможно за счет использования двудольного графа (пользователь-автор), весовой функции при определении возможных авторов на первом шаге алгоритма, и алгоритмов выделения устойчивых сочетаний, например, Brainsterm [3].

СПИСОК ЛИТЕРАТУРЫ

1. *Афонин С. А.*, *Гаспарянц А. Э.* Автоматическое построение функции оценки качества в задаче разрешения неоднозначности имен авторов научных публикаций. Программная инженерия, 2015, № 10, с. 31–37.
2. *Васенин В. А.*, *Афонин С. А.*, *Козицын А. С.* Автоматизированная система тематического анализа информации. — Информационные технологии, 2009, № 4, с. 1–32.

3. *Голомазов Д. Д.* Выделение терминов из коллекции текстов с заданным тематическим делением. — Информационные технологии, 2010, № 2, с. 8–13.