ОТЗЫВ

на автореферат диссертации А.С. Мохова «Метод классификации библиографической информации на основе комбинированных профилей классов с учетом структуры документов», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.01 - системный анализ, управление и обработка информации (в науке и промышленности)

В современных условиях быстрого роста объемов документальной информации одним из наиболее востребованных на практике направлений научных исследований становятся работы по обработке и анализу текстовых данных. Однако, несмотря на то, что эта проблематика активно развивается и находится в центре внимания целого ряда научных коллективов, тем не менее по многим важным вопросам до сих пор не найдено удовлетворительных ответов и не все предлагаемые решения соответствуют требованиям практики. В связи с этим актуальной представляется диссертация Андрея Сергеевича Мохова, в которой рассматриваются способы повышения точности классификации библиографической текстовой информации.

На основе экспериментальных исследований автором установлено, что профильные методы, использующие различные принципы отбора информативных терминов для построения профилей классов, классифицируют двуязычные библиографические текстовые документы с более высокой точностью, чем ряд известных методов (включая метод к-ближайших соседей, байесовский классификатор, метод опорных векторов и метод центроидов). Комбинируя различные способы построения профилей (статистический, теоретико-информационный и эвристический), автор разрабатывает новую группу алгоритмов, которые показывают более высокую точность. Наряду с комбинированными профилями для увеличения точности классификации соискатель применяет еще два эффективных прие-

ма: учитывает структуру библиографического описания научного документа (название, аннотация, ключевые слова) и строит коллективы решающих правил на основе разработанных ранее профильных алгоритмов и известных классификаторов.

Несмотря на логичное и последовательное изложение материала в автореферате, тем не менее имеется ряд замечаний.

1. В публикациях в области теории классификации большое внимание уделяется проблеме несбалансированных выборок (imbalanced samples). Такие выборки, в частности, характерны для классификации научных публикаций по заданным тематикам пользователя. К сожалению, в автореферате описание результатов экспериментов на несбалансированных выборках занимает один небольшой параграф. Не ясно насколько ухудшается точность распределения документов в малые классы при использовании авторских алгоритмов.
2. Не ясно, исследовалось ли влияние на точность таких факторов, как синонимия, наличие ключевых слов, пересекаемость классов (тематик, заданных пользователем).

Приведенные выше замечания не носят принципиального характера и не снижают ценности данного исследования. Судя по автореферату, диссертация А.С.Мохова представляет собой завершенную научно-квалификационную работу, выполненную на высоком профессиональном уровне, в ней содержатся новые теоретические результаты, позволяющие эффективно решать ряд практических задач классификации библиографических текстовых документов. На мой взгляд, данный научный труд удовлетворяет требованиям ВАК, предъявляемым к кандидатским диссертациям, и А.С.Мохов заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.01 - системный анализ, управление и обработка информации (в науке и промышленности).

Директор
Лаборатории информационных технологий ОИЯИ
доктор технических наук Кореньков В.В.

141980, Московская область,
г. Дубна, ул. Жолио-Кюри, 6
+7 (49621) 6-40-40; www.jinr.ru

Кореньков
Владимир Васильевич