«УТВЕРЖДАЮ»
Зам. директора по науке Федерального государственного бюджетного учреждения науки Институт проблем управления им. В.А.Трапезникова РАН
доктор физико-математических наук
_____ М.В. Губко
«29» мая 2017 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертационную работу Мохова Андрея Сергеевича «Метод классификации библиографической информации на основе комбинированных профилей классов с учетом структуры документов», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.01 – «Системный анализ, управление и обработка информации (в науке и промышленности по техническим наукам)»

1. Актуальность темы диссертационного исследования.

В условиях экспоненциального роста объёмов текстовой информации в Интернет, возрастают требования к методам классификации данных, представленных в виде документов определенного вида, например, научных публикаций или их библиографических характеристик. В настоящее время основное внимание исследователей направлено на создание новых и совершенствование существующих методов классификации для решения конкретных типов прикладных задач, поскольку именно особенности решаемой задачи определяют выбор эффективных классификаторов и методов предварительной обработки используемых текстовых данных.

В диссертации А.С.Мохова рассматривается специфический класс систем - персонализированные системы обработки и анализа библиографических научных публикаций. Такие системы позволяют проводить классификацию документов по заданным тематическим рубрикам пользователей (специалисты-предметники, научные сотрудники, аспиранты, преподаватели и др.) и предназначены для качественного улучшения их информационного обеспечения. При автоматизированном отборе и классификации текстовых документов, усиливаются требования к эффективности применяемых классификаторов, которые обеспечивали бы пользователя качественной информацией, необходимой для успешного проведения исследований.

Для повышения эффективности классификации двуязычных библиографических документов А.С.Мохов строит и комбинирует профили классов, состоящие из наиболее информативных русских и английских терминов, а также учитывает структуру библиографических документов

(название – аннотация - ключевые слова). Эффективность разрабатываемых классификаторов сравнивается с эффективностью таких хорошо известных в научной литературе методов, как метод *k*-ближайших соседей, наивный байесовский классификатор, метод опорных векторов, а также метод центроидов. Экспериментальные исследования, проведенные А.С.Моховым, подтверждают авторское предположение о том, что классификатор, который одновременно учитывает структуру документа и терминологическую информацию, содержащуюся в комбинированных профилях, способен увеличивать эффективность классификации русско-английских библиографических документов по сравнению с известными подходами.

Вышесказанное позволяет считать, что диссертационная работа А.С.Мохова, безусловно, является актуальной, поскольку посвящена решению важнейшей задачи классификационного анализа – повышению эффективности классификации текстовых документов.

2. Оценка научной новизны основных результатов исследования, выводов и рекомендаций, сформулированных в диссертации.

Научная новизна основных результатов диссертации заключается в том, что:

1. Разработана комплексная систематизация методов классификации многоязычной информации, на базе которой выявлены подходы, которые способны повысить эффективность классификации двуязычных библиографических документов.

2. Разработаны новые алгоритмы UNI, предназначенные для компенсации ограничений известных профильных методов, основанных на статистическом, теоретико-информационном и эвристическом способах отбора информативных терминов в профили классов. Расчет весов терминов в UNI проводится с помощью комбинирования трех этих подходов.

3. Разработан метод Struct, позволяющий увеличить точность классификации благодаря использованию комбинированных профилей и учету структуры библиографических документов. Предложено два способа формирования словарей для расчета весов терминов. Даны рекомендации по настройке параметров. Приведена оценка вычислительной сложности метода.

4. Сформированы и исследованы коллективы решающих правил (КРП), которые состоят преимущественно из новых методов, предложенных в работе. Сформированный КРП, состоящий из 5 классификаторов, увеличивает точность классификации двуязычных библиографических документов более чем на 5% по сравнению с наиболее точным из известных методов - PO-профилем.

5. Создана методика использования разработанных в диссертации программно-алгоритмических средств для построения высокоточных классификаторов с помощью предложенных в работе индивидуальных и коллективных методов классификации.

3. Оценка обоснованности и достоверности научных положений и выводов диссертации.

Достоверность и обоснованность научных положений, рекомендаций и выводов диссертации подтверждается корректным использованием методов математической статистики, теории вероятностей, системного анализа, а также результатами проведенных экспериментальных исследований на различных русско-английских выборках библиографических текстовых документов; сопоставлением собственных результатов с результатами известных работ; применением разработанных программно-алгоритмических средств при решении прикладных задач. Сопоставление прикладных результатов показывает их согласованность с теоретическими выводами и соответствует представлениям отечественных и зарубежных специалистов. Полученные автором результаты прошли апробацию на многих международных и всероссийских научно-технических конференциях и семинарах.

4. Значимость полученных автором диссертации результатов для науки и практики, рекомендации по их применению.

Важной особенностью работы является ее прикладная направленность на решение конкретной задачи – увеличение точности классификации в персонализированных системах обработки и анализа двуязычных библиографических научных публикаций. Автором создан программный комплекс (ПК) TextCat, в котором, наряду с известными классификаторами, реализованы методы классификации, разработанные в диссертации. Этот ПК позволяет проводить мониторинг научных журналов, скачивание и фильтрацию статей в соответствии с информационной потребностью пользователя. ПК TextCat предназначен для широкого круга исследователей, не имеющих специальных знаний в области программирования и теории классификации, он может адаптироваться к различным предметным областям и требованиям пользователя, в нем предусмотрена возможность наращивания функциональных возможностей путем включения новых модулей.

ПК TextCat внедрен в Институте проблем химической физики РАН. С его использованием проведен анализ специализированной базы данных, в которой хранится информация о научных публикациях сотрудников института. ПК TextCat внедрен в учебный процесс НИУ «Московский энергетический институт». На его основе разработаны четыре лабораторные работы по курсу «Интеллектуальные информационные системы». Применение ПК TextCat на практике подтверждается двумя актами о внедрении. Автором также получено свидетельство о регистрации программы для ЭВМ.

Для полномасштабного практического использования результатов диссертационной работы А.С.Мохова представляется целесообразно ознакомить с её основными положениями следующие организации, специализирующиеся в проведении исследований в области обработки и анализа документальной информации: Всероссийский институт научной и технической информации РАН (ВИНИТИ РАН), Институт научной информации по общественным наукам РАН (ИНИОН РАН), Всероссийский институт межотраслевой информации (ВИМИ), а также крупные академические и университетские центры, проводящие исследования в области

Data Mining, в том числе: Федеральный Исследовательский Центр «Информатика и управление» РАН, Институт проблем управления им. В.А.Трапезникова РАН, Институт проблем передачи информации им. А.А.Харкевича РАН, Институт системного программирования РАН, Научно-исследовательский вычислительный центр МГУ, МФТИ (ГУ), НИУ «Высшая школа экономики».

5. Содержание работы.

Основные результаты диссертации последовательно излагаются в следующих четырех главах.

В **первой главе** на основе системного анализа проводится декомпозиция процесса обработки текстовой информации, рассматриваются основные этапы этого процесса, особое внимание уделяется моделям представления документов и способам предварительной обработки данных, а также приводится систематизация методов классификации многоязычной информации и уточняется постановка задачи.

Во **второй главе** подробно анализируются профильные методы классификации, выявляются их тенденции к завышению весов терминов с высокой, средней и низкой частотой появления в выборке. С учетом специфики двуязычных документов для увеличения точности классификации на основе базовых профильных методов сформировано семейство новых алгоритмов UNI, в которых строятся комбинированные профили классов на основе совместного применения статистического, теоретико-информационного и эвристического подходов к взвешиванию терминов. Рассмотрены известные методы обработки и анализа текстовых данных, использующие информацию о структуре документов. Проведена декомпозиция библиографического документа на три раздела, которые упорядочены по важности с точки зрения увеличения точности классификации: название, ключевые слова и аннотация.

В **третьей главе** разработан новый метод Struct, учитывающий структуру библиографических документов. Проанализированы способы расчета весов терминов в методе Struct в зависимости от использования общего словаря или 3 отдельных словарей, построенных для терминов из названий, ключевых слов и аннотаций. Приводится описание выборок, используемых в исследованиях. Экспериментально обосновывается более высокая точность классификации в случае использования разработанных алгоритмов UNI5, UNI6 и метода Struct. Сформированы и исследованы коллективы решающих правил, которые состоят как из известных, так и разработанных автором классификаторов.

В **четвертой главе** описывается разработанный в диссертации программный комплекс TextCat и его применение при решении прикладных задач. Приводится методика использования ПК TextCat для построения высокоточных классификаторов путем применения индивидуальных методов классификации и формирования коллективов решающих правил.

6. Полнота публикаций научных результатов.

Результаты диссертации опубликованы в 3 научных изданиях из перечня ВАК Минобрнауки РФ, а также в 11 публикациях в трудах и материалах 9

общероссийских и международных конференций. Результаты диссертации обсуждались на многочисленных общероссийских и международных научных конференциях и семинарах. По результатам диссертационного исследования автором получено свидетельство о государственной регистрации электронных ресурсов (программа для ЭВМ - программный комплекс TextCat).

Основное содержание диссертационной работы достаточно полно отражено в автореферате и публикациях автора.

7. Замечания по диссертации.

Представленная работа А.С.Мохова не лишена недостатков.

1. В специализированной литературе по Data Mining достаточно часто используются статистический подход (хи-квадрат критерий) и теоретико-информационный подход (например, критерий взаимной информации) для выявления информационных признаков в задачах обработки фактографических и документальных данных. Автор применяет те же критерии для построения профилей. При этом в работе не уделяется достаточного внимания анализу различий в использовании этих подходов для выявления информативных признаков и построения профилей.
2. При анализе результатов поиска и классификации текстовой информации широкое применение получили показатель «точность-полнота» и F-мера, которые позволяют более детально проанализировать, какие из документов ошибочно отнесены к другим классам. Было бы целесообразно их использовать при проведении исследований, описанных в главе 3.
3. Выборки из базы данных Института проблем химической физики РАН (глава 4) содержат только названия и не позволяют применить авторский метод Struct, который предназначен для анализа библиографических документов. Не совсем понятно, зачем автор использовал такую «ущербную» базу данных как прикладную задачу.
4. Автором сформированы достаточно разнотипные выборки, на которых проводились исследования. Причем, несмотря на одинаковую структуру выборок, получен достаточно большой разброс ошибок. Можно ли сделать какие-то обобщающие выводы о том, какие выборки и почему приводят к ухудшению точности рассматриваемых методов, - это очень важный элемент сравнительного анализа.
5. В настоящее время в области Data Mining доступно для использования свободно распространяемое программное обеспечение, апробированное на известных документальных коллекциях. Не ясно, применялось ли оно в работе для обработки текстовых документов.

8. Общая оценка диссертационной работы.

Отмеченные выше недостатки не влияют на положительную оценку диссертационной работы А.С.Мохова в целом.

Диссертация А.С.Мохова представляет собой законченное научное исследование, выполненное лично автором на актуальную тему и на высоком научном уровне.

Диссертация и автореферат написаны ясным языком с использованием общепринятой в области системного анализа, управления и информационных технологий терминологии. Работа хорошо структурирована по содержанию и отличается логичностью изложения материала. Все основные положения и выводы достаточно полно аргументированы. Содержание автореферата отражает основные результаты работы и соответствует содержанию диссертации. Оформление диссертации и автореферата соответствует требованиям ВАК. Основные результаты диссертационной работы получены лично соискателем.

Отзыв обсужден и одобрен на научном межлабораторном семинаре Федерального государственного бюджетного учреждения науки Институт проблем управления им. В. А. Трапезникова РАН (ИПУ РАН) 29 мая 2017 года, протокол № 2. Председатель семинара - д.т.н., проф. Дорофеюк Александр Александрович, секретарь семинара – к.т.н., доц. Чернявский Александр Леонидович.

9. Заключение.

Все вышеизложенное позволяет сделать вывод, что диссертационная работа «Метод классификации библиографической информации на основе комбинированных профилей классов с учетом структуры документов» отвечает всем требованиям «Положения о присуждении ученых степеней» (в редакции Постановления Правительства Российской Федерации от 24 сентября 2013 г. № 842), предъявляемым к кандидатским диссертациям, а ее автор, Мохов Андрей Сергеевич, заслуживает присуждения ему ученой степени кандидата технических наук по специальности 05.13.01 – «Управление в социальных и экономических системах (технические науки)».

Председатель семинара:
доктор технических наук, профессор,
главный научный сотрудник ИПУ РАН
Дорофеюк Александр Александрович
лаборатория №55 „Обработка больших массивов информации в иерархических системах"

(подпись)

Сведения о составителях отзыва:

Фамилия, имя, отчество: Дорофеюк Александр Александрович
Ученая степень: доктор технических наук
Ученое звание: профессор
Место работы: Федеральное государственное бюджетное учреждение науки
Институт проблем управления им. В.А. Трапезникова РАН (ИПУ РАН)
Должность: главный научный сотрудник
Почтовый адрес: 117997, г. Москва, ул. Профсоюзная, д. 65
Телефон: +7 (495) 334-75-40
E-mail: daa2@mail.ru