«Утверждаю»
Проректор по научной деятельности
Казанского (Приволжского) федерального университета,
профессор

Д.К. Нургалиев
2 сентября 2014 года

ОТЗЫВ

ведущей организации на диссертационную работу А.А. Алексеева «Метод автоматического аннотирования новостных кластеров на основе тематического анализа», представленную на соискание ученой степени кандидата физико-математических наук по специальности «05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» в диссертационный совет Д 501.001.44 в Московском государственном университете имени М.В. Ломоносова

Актуальность темы

В диссертационной работе А.А. Алексеева исследуется задача выявления различных вариантов упоминаний основных участников новостного сюжета, описываемого набором тематически близких новостных документов. В качестве целевой задачи выступает задача автоматического аннотирования новостных кластеров, решение которой во многом зависит от полноты информации о внутреннем устройстве исходной коллекции для аннотирования, в частности, наличия информации о вариативности именования сущностей. Подобная информация необходима для объективной оценки значимости участников новостного кластера и, соответственно, их правильного представления в автоматической аннотации. Несмотря на широкий спектр приложений, таких, как коммерческие новостные сервисы, специализированные инструменты по мониторингу средств массовой информации и т. д., задача автоматического аннотирования новостных кластеров еще далека от своего полного решения, поэтому тема диссертационной работы несомненно является весьма актуальной.

Основные положения и результаты работы

Диссертационная работа состоит из введения, 4 глав, заключения, списка литературы, содержащего 82 источника, и двух приложений. Общий объем работы составляет 122 страницы.

Во введении полно и подробно изложены история изучаемых вопросов, мотивация исследования, актуальность темы диссертационной работы, методика исследования, а также основные результаты диссертации.

В первой главе приведен обзор подходов к решению задачи автоматического аннотирования, в рамках которого раскрыта проблематика проведенного исследования применительно к различным областям и типам коллекций для аннотирования. Отмечено, что построение автоматической аннотации в значительной степени связано со спецификой аннотируемой коллекции, а также зависит от требуемого типа аннотации. Рассмотрены основные виды коллекций для аннотирования, описаны существующие методы автоматического аннотирования и модели представления данных в контексте предлагаемой классификации. При этом особое внимание уделено именно моделям представления данных, в частности, способам учета лексико-семантической вариативности в существующих подходах и методам автоматического аннотирования. В результате проведенного анализа предметной области диссертационного исследования сделано заключение о том, что большинство существующих методов автоматического аннотирования работает на основе пословной модели представления – представления входной коллекции текстовых документов в виде массива отдельных слов, что влечет за собой отсутствие инструментов для учета вариативности естественного языка. Данный факт лежит в основе проведенного исследования. Также в первой главе представлено описание существующих автоматических и полуавтоматических подходов к оценке и сравнению качества автоматических аннотаций.

Вторая глава диссертации посвящена различным моделям представления текстовых коллекций при автоматической обработке и, в особенности, отражению вариативности именования сущностей в данных моделях. Предложена новая модель тематических цепочек новостного кластера, в рамках которой все варианты упоминаний некоторого участника новостного сюжета объединяются в единую структуру, называемую тематической цепочкой. Важной особенностью данной модели является одинаковый учет различных вариантов упоминаний как с помощью отдельных слов, так и с помощью многословных выражений. В этой же главе предложен алгоритм автоматического построения тематических цепочек новостного кластера, основанный на шести характеристиках схожести различной природы: контекстно-зависимые и контекстно-независимые признаки, учитывающие статистические особенности рассматриваемой коллекции и абсолютные знания о наблюдаемых сущностях соответ-

ственно. Все характеристики схожести единообразно учитываются в процессе общего ранжирования пар языковых выражений.

В третьей главе исследованы вопросы интеграции предложенной модели тематических цепочек в методы автоматического аннотирования и аннотирования на основе комплексных моделей представления в целом. Отдельно отметим предложенный диссертантом универсальный алгоритм расширения пословной модели представления новыми знаниями, заложенными в модели тематических цепочек. Подобная интеграция теоретически должна улучшать качество существующих методов автоматического аннотирования, работающих на основе пословной модели. Эта гипотеза подтверждена полученным улучшением качества популярных методов автоматического аннотирования MMR и Sumbasic по всем примененным способам оценки. Полученный результат в полной мере подтверждает значимость предложенной модели тематических цепочек для задачи автоматического аннотирования.

Четвертая глава посвящена описанию разработанного программного комплекса, в котором имплементированы все предложенные модели и алгоритмы.

Основными научными результатами диссертации являются:

- метод автоматического построения модели основных участников новостного кластера, основанный на комбинировании разнородных признаков сходства;
- применение построенной модели в существующих методах автоматического аннотирования и создание на ее основе двух новых методов автоматического аннотирования.

Названные результаты являются **новыми и достоверными**. Разработанная модель не зависит от предметной области и может применяться в различных задачах автоматической обработки новостных кластеров.

Практическая значимость полученных результатов заключается в том, что на основе предложенного алгоритма спроектирована и реализована многомодульная программная система, прошедшая практическую апробацию.

Материал диссертации изложен ясным языком, на хорошем уровне формализации, со строгими определениями и четкими математическими формулировками. Использован современный математический аппарат.

Замечания

Диссертационная работа А.А. Алексеева не лишена недостатков. Большая их часть относится к оформлению результатов:

1) На стр. 20 приведен рис. 1, описание которого в тексте отсутствует, кроме того, на этот рисунок нет ссылок в тексте;
2) Имеются недостатки в оформлении формул, в частности, часто в одной и той же формуле использованы различные шрифты для обозначения одних и тех же величин (для обозначения документов (стр. 22), вероятностей (стр. 82), максимума и минимума (по всему тексту) и др.); без пояснений используется обозначение arg max; косинусная мера (стр. 28) введена без ссылок, поэтому о введенных обозначениях в ней приходится только догадываться (нормы, скалярное произведение); не ясна роль обозначения с косинусом, содержащим необъясненную величину в скобках; на стр. 42 при описании качества контента использован без пояснений нормализующий фактор Z;
3) Модифицированная мера Жаккара, введенная на стр. 59, содержит нормализующий коэффициент, значение которого, без всяких пояснений и мотивации, выбрано равным 3;
4) На стр. 73 сформулирована и далее доказана лемма 1. Других строгих утверждений диссертация не содержит. Как принято в математике, леммой называют вспомогательное утверждение, используемое далее для доказательства теоремы. В данном случае использование названия «лемма» представляется нецелесообразным.

В качестве еще одного замечания отметим, что, по нашему мнению, описание собственных результатов могло бы быть более подробным. Так, например, на стр. 50 автор ссылается на собственную статью [3], говоря о проведенных экспериментах. Одновременно обзорные части диссертации значительно превышают половину объема работы.

Заключение

Отмеченные недостатки не снижают в целом положительную оценку диссертационного исследования А.А. Алексеева. Диссертация является завершенной научно-квалифицированной работой, которая обладает необходимой степенью обоснованности научных положений, выводов и заключений, достоверностью и новизной полученных результатов. Задача выявления лексико-семантической вариативности имеет важное теоретическое значение, а ее решение, предложенное в диссертации, представляет высокую научную ценность.

Диссертационная работа соответствует паспорту специальности «05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей». Автореферат достаточно полно и объективно отражает содержание диссертации.

Основные результаты опубликованы в 14 печатных работах, из них 3 – в периодических изданиях, входящих в перечень ВАК РФ, 3 проиндексированы в БД Скопус. Основные результаты диссертационного исследования доложены на международных и всероссийских научных конференциях.

Диссертационная работа удовлетворяет требованиям Положения о порядке присуждения ученых степеней, утвержденного Постановлением Правительства РФ, предъявляемым к кандидатским диссертациям по специальности «05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей», а ее автор, Алексеев Алексей Александрович, заслуживает присуждения искомой ученой степени кандидата физико-математических наук.

Отзыв на диссертацию обсужден и одобрен на семинаре Отдела высокопроизводительных вычислений и систем НИЦ «НИИ математики и механики им. Н.Г. Чеботарева» Института математики и механики имени Н.И. Лобачевского Казанского (Приволжского) федерального университета (протокол № 5 от 28.08.2014 г.).

Доктор физико-математических наук, профессор, заместитель директора по научной деятельности Института математики и механики имени Н.И. Лобачевского Казанского (Приволжского) федерального университета, Заслуженный деятель науки Республики Татарстан

А.М. Елизаров

Кандидат физико-математических наук, доцент кафедры теории функций и приближений Института математики и механики имени Н.И. Лобачевского Казанского (Приволжского) федерального университета

Е.К. Липачёв

Почтовый адрес: 420008, г. Казань, ул. Профессора Нужина, д. 1/17
Электронный адрес: amelizarov@gmail.com
Телефон: +7 (987) 2961781