ОТЗЫВ

официального оппонента о диссертации Алексея Александровича Алексеева «Метод автоматического аннотирования новостных кластеров на основе тематического анализа», представленной на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Актуальной задачей обработки новостных потоков является автоматическое аннотирование новостных кластеров – выделение наиболее важной информации из набора новостных документов, посвященных одному событию или набору связанных событий. Подобная обработка широко используется для анализа средств массовой информации и является неотъемлемой частью специализированных аналитических систем. Качественное решение задачи автоматического аннотирования требует глубокого понимания анализируемой коллекции, особенно важным является установление ключевых участников ситуации, необходимое для объективного определения их значимости и значимости взаимодействий между ними.

Известно, что тексты на естественном языке обладают свойством лексико-семантической вариативности, разрешение которой необходимо для определения наиболее значимых участников новостного сюжета и, соответственно, для решения актуальной задачи автоматического аннотирования новостных кластеров. Именно задаче разрешения лексико-семантической вариативности в новостных кластерах и посвящена кандидатская диссертация А.А. Алексеева.

Диссертация состоит из введения, четырех глав, заключения, списка литературы из 82 источников и двух приложений. Объем работы составляет 122 страницы.

Во введении обосновывается актуальность темы диссертационной работы, кратко описываются основные результаты диссертации, методика исследования, основные принципы предлагаемого подхода и даются сведения об апробации результатов.

Первая глава посвящена задаче автоматического аннотирования, описываются различные типы аннотаций, которые выстраиваются в единую классификацию. Для каждого типа аннотаций дается описание областей применения – как текущих, так и перспективных. Кроме того, в данной главе представлены основные идеи и принципы алгоритмов подготовки автоматических аннотаций, при этом указываются соответствующие модели представления данных. Особое внимание уделяется учету лексико-семантической вариативности в существующих моделях и методах автоматического аннотирования, проводится сравнительный анализ преимуществ и недостатков используемых моделей представления в контексте вариативности естественного языка. Кроме того, в первой главе

подробно описываются методы оценки качества порожденных аннотаций, а также методы сравнения автоматических аннотаций между собой.

Во второй главе вводится понятие тематической цепочки, которая является формальным представлением участника новостного сюжета с учетом всех вариантов его именования внутри рассматриваемого новостного кластера. Данный подход позволяет единообразно учитывать различные способы упоминаний участников новостного сюжета, которые могут использоваться в текстах на естественном языке – как с помощью отдельных слов, так и с помощью словосочетаний. Кроме того, во второй главе представлено описание алгоритма автоматического наполнения предложенной модели тематических цепочек. Данный алгоритм основан на шести разнородных характеристиках схожести, которые подразделяются на два типа: контекстно-зависимые характеристики, учитывающие статистическую информацию из рассматриваемого новостного кластера, и контекстно-независимые, отражающие знания о близости языковых выражений из вспомогательных источников. Алгоритм расчета и значимость каждой из характеристик схожести подробно рассмотрены в тексте диссертации. Веса всех характеристик схожести нормированы и объединены в единый алгоритм ранжирования языковых выражений, что делает метод построения тематических цепочек универсальным и масштабируемым в отношении добавления новых характеристик схожести. Одним из важных результатов главы является математическое исследование меры схожести между предложениями в случае усложнения пословного представления с помощью применения нескольких операций.

В третьей главе разрабатываются два новых метода автоматического аннотирования на основе предложенной модели тематических цепочек. Первый метод автоматического аннотирования позволяет учитывать упоминания тематических цепочек в предложениях текстов, а второй метод - существование связей между тематическими цепочками. Также предлагается универсальный алгоритм обогащения пословной модели представления дополнительной информацией, содержащейся в модели тематических цепочек. Необходимо отметить, что пословная модель представления используется большинством наиболее распространенных методов автоматического аннотирования, а улучшение качества базовой модели позволит улучшить качество и самих методов аннотирования. Подобный эксперимент был проведен и представлен в третьей главе. Сравнение автоматических аннотаций на основе пословной и обогащенной тематическими цепочками моделей представления производится с помощью двух различных способов оценки: автоматических мер качества ROUGE и полуавтоматического метода «Пирамиды». Полученные результаты подтверждают, что модификации методов на основе обогащенной модели демонстрируют лучшее качество по сравнению с базовой моделью по всем способам оценки. Данный факт доказывает ценность разработанной модели, а также предложенного алгоритма обогащения пословной модели представления дополнительной информацией.

В четвертой главе описывается программный комплекс, который был разработан в рамках диссертационного исследования. Данный комплекс реализует предложенный метод

наполнения модели тематических цепочек, а также все используемые в работе методы автоматического аннотирования. Кроме того, программный комплекс содержит модуль для оценки автоматических аннотаций по методике ROUGE, что, в совокупности с остальными модулями, делает его универсальным инструментом для исследований в данной области и решения прикладных задач.

По работе имеются следующие замечания:

1) Проведенное во второй главе исследование меры схожести между предложениями в случае усложнения пословного представления важно для диссертации по физико-математическим наукам, однако этот результат не отражен в выводах ко второй главе (раздел 2.7). Этот результат не указан и в заключении к диссертационной работе.

2) В разделе 2.3. (стр. 48) говорится, что под темой/подтемой понимается предикат $P(C_1, ..., C_n)$, и его атрибуты C_1 , ..., C_n называются тематическими элементами. С позиций устоявшейся терминологии логики предикатов первого порядка корректно было бы говорить об атомарной формуле вида $P(C_1, ..., C_n)$, где P – имя n–арного предиката.

3) В разделе 2.3. (стр. 48) вместо названия свойства бинарных отношений «рефлексивность» использовано название «рефлективность».

4) В разделе 2.3. (стр. 49) вводятся без пояснения однонаправленная и двунаправленная стрелки, используемые для отображения раскрытия темы через тематические элементы.

5) В подразделе 2.4.1 (стр. 53) дано некорректное обозначение множества слов w.

Отмеченные недостатки не снижают общего высокого научного уровня диссертации, не подвергают сомнению достоверность и научную обоснованность всех полученных автором результатов и не влияют на общую положительную оценку диссертационной работы.

Результаты диссертационного исследования полностью опубликованы в 14 статьях, в том числе три работы опубликованы в изданиях из перечня ВАК, прошли апробацию на научных конференциях и семинарах. Автореферат правильно и полно отражает содержание диссертации.

Таким образом, диссертация представляет собой законченное научное исследование по актуальной теме. В работе получены новые результаты, имеющие несомненную научную значимость. Диссертационная работа «Метод автоматического аннотирования новостных кластеров на основе тематического анализа» удовлетворяет всем требованиям ВАК РФ, предъявляемым к кандидатским диссертациям, а ее автор, Алексеев Алексей Александрович, заслуживает присвоения ученой степени кандидата физико-математических наук по

специальности 05.13.11 – "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей".

Профессор, доктор технических наук,
профессор факультета бизнес-информатики
НИУ ВШЭ

_____ В.А. Фомичев

Адрес: НИУ ВШЭ, факультет бизнес-информатики,
Кирпичная ул. 33, 105187 Москва
Электронная почта: vfomichov@hse.ru
Моб. тел.: +7-903-505-44-51

25.08.2014

ПОДПИСЬ ЗАВЕРЯЮ