

В Диссертационный совет Д.002.199.01
при Федеральном государственном
бюджетном учреждении науки
Санкт-Петербургский институт
информатики и автоматизации
Российской академии наук
199178, Санкт-Петербург, 14 линия, д. 39

Отзыв на автореферат диссертации

Карповича Сергея Николаевича

На тему «*Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов*», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.11 – «*Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей*».

В работе Карповича С.Н. рассматривается задача построения вероятностной тематической модели для потока текстовых документов. Данный тип моделей используется для классификации текстовых документов на принадлежность к определенной теме или категории. Разработка и исследование таких моделей необходимы для развития соответствующих цифровых технологий. В условиях постоянно увеличивающегося объема информации в сети Интернет и, в особенности, роста количества неструктурированных текстовых документов, тему диссертационного исследования Карповича С.Н. можно оценить как весьма актуальную.

Обоснованность научных положений, вынесенных на защиту, обусловлена применением известных методов исследования, результатами выполненного анализа научных работ по тематике диссертации, подтверждением используемых теоретических предположений результатами проведенного вычислительного эксперимента. Уровень строгости изложения соответствует общепринятым в технических специальностях научным стандартам. Новизна и значимость научно-технических результатов подтверждается полученными публикациями и свидетельством на программу для ЭВМ.

Новизна полученных автором результатов, заключается в возможности использования предложенных методов и алгоритмов для построения вероятностной тематической модели потока текстовых документов, что ранее не достигалось существующими методами. Для обеспечения этой возможности автором были улучшены методы расчета матриц ВТМ на основе обучения с учителем, предложен алгоритм многозначной классификации текстовых документов, полученных из потока документов, и использовано произведение Адамара для определения тем нового слова. В совокупности, предложенные методы позволяют динамически перестраивать вероятностную тематическую модель при появлении новых документов.

Результаты, представленные в диссертационной работе, обладают теоретической значимостью. В частности, использование обучения с учителем для расчета матриц вероятностной тематической модели, позволяет ограничиться единственной итерацией при расчете. Предложенный автором алгоритм многозначной классификации текстовых документов позволяет составлять динамические тематические модели, предназначенные для анализа поступающего потока текстовых документов. Предлагаемое использование произведения Адамара для определения вектора тем слов, ранее не встречавшихся в модели, позволяет динамически перестраивать составленную вероятностную тематическую модель.

Значимость результатов для практического использования заключается в разработке русскоязычного корпуса текстовых документов, содержащего не только сами документы,

но и метаинформацию, такую как автор документа, время описанных событий, тема. Данный корпус распространяется по свободной лицензии, может быть модифицирован и использован для задач вероятностного тематического моделирования. Также автором был разработан комплекс программных средств для анализа потока текстовых документов, основанный на использовании микросервисной архитектуры. Предполагается, что такие микросервисы могут быть адаптированы для разработки других программных средств для тематического моделирования, а полученные средства визуализации позволят наглядно изобразить изменение интереса к определенным темам с течением времени. Эти научные технические результаты были внедрены в системе анализа новостного потока для Правительства Москвы и для многозначной классификации поисковых запросов в «ООО Rambler&Co».

В качестве замечаний необходимо отметить следующее.

1. В тексте автореферата не представлена система параметров, характеризующая размеры и другие свойства получаемого корпуса текстов, имеющие значимое влияние на сложность предлагаемых алгоритмов для автоматической обработки текстов.
2. В тексте автореферата не приведены результаты по теоретическому исследованию (например, в виде теорем) таких типовых свойств предлагаемых алгоритмов, как корректность, точность, времененная и емкостная сложность.
3. В тексте автореферата не приведены сводные числовые характеристики полученного комплекса программных средств, характеризующие объем выполненной работы по разработке программного обеспечения.

Указанные замечания не снижают общего научного уровня рецензируемой работы. В целом, считаю, что содержание автореферата свидетельствует о том, что диссертация представляет собой законченное научное исследование, результаты которого обладают научной новизной. Работа Карповича Сергея Николаевича отвечает требованиям, установленным п. 9 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 № 842, предъявляемым к кандидатским диссертациям. Считаю, что автор диссертационной работы заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Отзыв составил:

Доцент ФГБОУ ВО «Петрозаводский государственный университет»,
к.ф.-м..н., доц.

Корзун Д.Ж.

Сведения о составителе отзыва:

Фамилия, имя, отчество: Корзун Дмитрий Жоржевич
Ученая степень: кандидат физико-математических наук
Ученое звание: доцент
Место работы: Федеральное государственное бюджетное образовательное учреждение высшего образования «Петрозаводский государственный университет»
Должность: доцент, ведущий научный сотрудник
Почтовый адрес: 185910, Республика Карелия, г. Петрозаводск, пр. Ленина, 33
телефон: +78142711084, +79095689571
адрес электронной почты: dkorzun@cs.karelia.ru

Подпись Корзун Д.Ж. заверена
специалистом по кадрам
06.10.2017