УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2016 Том 16 № 3 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2016 Vol. 16 No 3 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.91

АВТОМАТИЧЕСКАЯ СУММАРИЗАЦИЯ ВЕБ-ФОРУМОВ КАК ИСТОЧНИКОВ ПРОФЕССИОНАЛЬНО ЗНАЧИМОЙ ИНФОРМАЦИИ

К.И. Бурая[a], П.Д. Виноградов[b,c], В.А. Грозин[b,d], Н.Ф. Гусарова[b], Н.В. Добренко[b], В.А. Трофимов[b,e]

[a] ЗАО «Петер-Сервис», Санкт-Петербург, 191123, Российская Федерация
[b] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[c] ООО ЛКМСПБГУ, Санкт-Петербург, 199034, Российская Федерация,
[d] Диджинетика, Москва, 125319, Российская Федерация
[e] Санкт-Петербургский Центр Разработок ЕМС, Санкт-Петербург, 199004, Российская Федерация
Адрес для переписки: natfed@list.ru

Информация о статье
Поступила в редакцию 16.03.16, принята к печати 10.04.16
doi: 10.17586/2226-1494-2016-16-3-482-496
Язык статьи – русский

Ссылка для цитирования: Бурая К.И., Виноградов П.Д., Грозин В.А., Гусарова Н.Ф., Добренко Н.В., Трофимов В.А. Автоматическая суммаризация веб-форумов как источников профессионально значимой информации // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 3. С. 482–496. doi: 10.17586/2226-1494-2016-16-3-482-496

Аннотация

Предмет исследования. Конкурентным преимуществом современного специалиста является максимально широкий охват источников информации, полезных с точки зрения получения и освоения актуальной профессионально значимой информации. Среди таких источников значительное место занимают профессиональные веб-форумы. В статье рассматривается задача автоматической суммаризации текста форума, т.е. выделения тех его фрагментов, которые содержат профессионально значимую информацию. **Метод.** Исследование строится на базе статистического анализа текстов форумов посредством машинного обучения. Для исследований отобраны шесть веб-форумов, тематикой которых являются аспекты технологий различных предметных областей. Разметка форумов проводилась экспертным путем. С использованием различных методов машинного обучения построены модели, отражающие функциональную связь между оцениваемыми характеристиками качества извлечения профессионально значимой информации и признаками постов. Для оценки качества моделей использованы кумулятивная метрика *NDCG* и ее дисперсия. **Основные результаты.** Показано, что в оценке эффективности извлечения профессионально значимой информации важную роль играет контекст запроса. Отобраны характерные для извлечения профессионально значимой информации контексты запросов, отражающие различные трактовки информационной потребности пользователей, обозначенные терминами релевантность и информативность. Построены шкалы для их оценок, соответствующие общемировым подходам. Экспериментально подтверждено, что результаты суммаризации форумов, выполняемой экспертами вручную, существенно зависят от контекста запроса. Показано, что в общей оценке эффективности извлечения профессионально значимой информации релевантность достаточно хорошо описывается линейной комбинацией признаков, а для оценки информативности уже требуется их нелинейная комбинация. При этом при оценке релевантности ведущую роль играют признаки, связанные с ключевыми словами, а при оценке информативности на первый план выступают характеристики текста поста в целом, а также признаки, связанные со структурой треда как текста и как социального графа. Показано, что эффективность извлечения информативных постов слабо зависит от способа задания ключевых слов, в то время как для извлечения релевантных постов такая зависимость существенна. Выявлен способ выделения ключевых слов, наиболее эффективный для реальных приложений. Показано, что при выделении релевантных постов линейные методы выигрывают в эффективности по сравнению с нелинейными, а модель LDA занимает промежуточное положение; в то же время при выделении информативных постов линейные и нелинейные методы идентичны по эффективности, а модель LDA значительно уступает им обоим. Предложена содержательная модель, позволяющая объяснить полученные результаты. **Практическая значимость.** Полученные результаты могут служить основой для построения и новых и адекватного применения существующих алгоритмов суммаризации веб-форумов, что позволит существенно сократить временные и ресурсные затраты пользователя на получение и изучение максимально свежей профессионально значимой информации.

Ключевые слова

профессионально значимая информация, суммаризация веб-форумов, релевантный пост, информативный пост, машинное обучение, классификационные модели, регрессионные модели, линейные методы, нелинейные методы, латентное размещение Дирихле, связность текста, социальный граф

AUTOMATIC SUMMARIZATION OF WEB FORUMS AS SOURCES OF PROFESSIONALLY SIGNIFICANT INFORMATION

K.I. Buraya[a], P.D.Vinogradov[b,c], V.A. Grozin[b,d], N.F. Gusarova[b], N.V. Dobrenko[b], V.A.Trofimov[b,e]

[a] Peter-Service, Saint Petersburg, 191123, Russian Federation
[b] ITMO University, Saint Petersburg, 197101, Russian Federation
[c] Computer Modelling Laboratory of the St. Petersburg State University (CML SPBGU), Saint Petersburg, 199034, Russian Federation
[d] Diginetica, Moscow, 125319, Russian Federation
[e] EMC Corporation, Saint Petersburg, 199004, Russian Federation
Corresponding author: natfed@list.ru

Article info
Received 16.03.16, accepted 10.04.16
doi: 10.17586/2226-1494-2016-16-3-482-496
Article in Russian
For citation: Buraya K.I., Vinogradov P.D., Grozin V.A., Gusarova N.F., Dobrenko N.V., Trofimov V.A. Automatic summarization of web forums as sources of professionally significant information. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 3, pp. 482–496. doi: 10.17586/2226-1494-2016-16-3-482-496

Abstract
Subject of Research. The competitive advantage of a modern specialist is the widest possible coverage of information sources useful from the point of view of obtaining and acquisition of relevant professionally significant information. Among these sources professional web forums occupy a significant place. The paper considers the problem of automatic forum text summarization, i.e. identification of those fragments that contain professionally relevant information. **Method.** The research is based on statistical analysis of texts of forums by means of machine learning. Six web forums were selected for research considering aspects of technologies of various subject domains as their subject-matter. The marking of forums was carried out by an expert way. Using various methods of machine learning the models were designed reflecting functional communication between the estimated characteristics of PSI extraction quality and signs of posts. The cumulative *NDCG* metrics and its dispersion were used for an assessment of quality of models. **Main Results.** We have shown that an important role in an assessment of PSI extraction efficiency is played by request context. The contexts of requests have been selected, characteristic of PSI extraction, reflecting various interpretations of information needs of users, designated by terms relevance and informational content. The scales for their estimates have been designed corresponding to worldwide approaches. We have experimentally confirmed that results of the summarization of forums carried out by experts manually significantly depend on request context. We have shown that in the general assessment of PSI extraction efficiency relevance is rather well described by a linear combination of features, and the informational content assessment already requires their nonlinear combination. At the same time at a relevance assessment the leading role is played by the features connected with keywords, and at an informational content assessment characteristics of the post text in general come to the fore, and also the features connected with structure of a thread as the text and the social graph. We have shown that efficiency of extraction of informative posts poorly depends on a way of keywords assignment while such dependence is essential to extraction of relevant posts. The way of keywords extraction, the most effective for real appendices has been revealed. We have shown that at extraction of relevant posts linear methods are better in efficiency in comparison with nonlinear, and the LDA model is intermediate; at the same time at extraction of informative posts linear and nonlinear methods are identical by efficiency, and the LDA model considerably concedes to both of them. We have proposed substantial model explaining the received results. **Practical Relevance.** The obtained results can provide background for creation of new and adequate application of the existing algorithms of web forums summarization that will allow reducing significantly user's time and resource expenditure by receiving and studying the last minute professionally significant information.

Keywords
professionally significant information, web forums summarization, relevant post, informative post, machine learning, classification models, regression models, linear methods, nonlinear methods, latent Dirichlet allocation, text connectivity, social graph

Введение

Непрерывное обновление профессиональных знаний – один из ключевых факторов успеха специалиста, определяющий его востребованность на рынке труда. Современная система профессионального образования реализует это требование через компетентностно-ориентированный подход [1–3].

Профессиональные знания – это характеризующие особенности конкретной деятельности сведения, которые необходимы для эффективной ее реализации [4]. В рамках компетентностного подхода чаще используются термины «профессионально значимая информация (ПЗИ), направленная на формирование профессиональных компетенций» [5] или «профессионально важная информация» [6]. В условиях динамически изменяющихся требований рынка конкурентными преимуществами специалиста становится максимально широкий охват источников информации, полезных с точки зрения получения и освоения актуальной ПЗИ. Среди таких источников все более значительное место занимают веб-форумы. Механизм работы веб-форума состоит в создании пользователем своей темы (треда) с его последующим обсуждением путем размещения другими пользователями (посетителями форума) сообщений (постов) внутри треда. Тем самым формируется генерируемый пользователями контент, представляющий собою определенным образом структурированный текст, который фиксирует процесс совместной работы (дискуссии)

пользователей по заявленной теме. Потенциальными источниками ПЗИ являются технические форумы, которые, как правило, характеризуются узкой проблемной направленностью и изложением в постах конкретного профессионального опыта.

В работах [7–9] выделены преимущества веб-форума как источника ПЗИ по сравнению с традиционными учебными и научными изданиями.

– Форум содержит максимально свежую и оперативную информацию по теме. Конкретные технологические решения часто формируются в ходе дискуссии, в то время как для их публикации в виде статьи и, тем более, учебного издания требуется большой срок.
– В постах форума отражен опыт людей, непосредственно использующих конкретную технологию, причем как позитивный, так и негативный опыт – такая информация практически недоступна в официальной документации.
– Информация на форуме представлена в структурированном виде, что расширяет возможности информационного поиска.
– Представление информации на форуме значительно более эффективно с точки зрения передачи профессионального опыта. Изложение информации на форуме отличается большей свободой стиля, наличием эмоциональных оценок, различными типами визуализации.
– Информация на форуме отражает коллективное мнение профессионального сообщества, что соответствует требованиям Закона об образовании.

В то же время при анализе форума как источника ПЗИ возникает ряд проблем [10]:

– информационная избыточность – большой объем повторяющейся, сугубо эмоциональной и профессионально несущественной информации;
– дрейф темы – постепенный переход от первоначально заявленной темы на другие;
– ограничение обсуждения формой «ответ на ответ», что снижает его полноту;
– недостатки языка постов – неполный формат предложений, различия в понимании смысла концептов в отдельных постах, что особенно затрудняет анализ форумов на иностранных языках.

Рассмотрим типичную ситуацию. Специалист узнает о недавно появившейся технологии, которая может быть перспективной в его профессиональной деятельности. Поисковый запрос приводит его на форум, где эта технология обсуждается. Достаточно ли на нем информации для оценки ее перспективности и для детального знакомства с ней? Какие именно посты содержат ПЗИ? Хотелось бы заранее получить ответы на эти вопросы и только потом детально изучать выделенные посты. Таким образом, при анализе веб-форума как источника ПЗИ возникает задача автоматической суммаризации текста форума, т.е. выделения тех его фрагментов, которые покрывают соответствующую ПЗИ. Актуальность задачи дополнительно возрастает, если форум ведется на незнакомом языке и доступен специалисту только через перевод – дорогостоящую и ресурсоемкую операцию. Предложены различные подходы к суммаризации текстовых форумов [11]. Наиболее перспективны с точки зрения обозначенной задачи методы машинного обучения [12–14]. Однако спектр таких методов весьма широк, и выбор наиболее эффективного метода для конкретной задачи суммаризации представляет проблему.

Существенную роль при суммаризации форумов играет процедура извлечения ключевых слов. Методы отбора ключевых слов делятся на две категории: назначение ключевых слов – отбор подходящих слов из предопределенного словаря или таксономии в соответствии с контентом документа, и извлечение ключевых слов, когда они отбираются непосредственно из анализируемого документа. Методы второй группы, в свою очередь, делятся на несколько категорий [15, 16] – машинное обучение (МО), лингвистические, графовые, статистические, эвристические методы. Статистические методы основаны на вычислении различных статистик слов в документе, в том числе частоты встречаемости слова, tf.idf, *n*-грамм и пр. [14]. Эвристические методы [17] используют такие характеристики, как позиция слова в документе, наличие элементов форматирования, длина фрагмента документа и т.п., т.е. позволяют в определенной степени учесть структуру документа. Однако проблемно-ориентированная оценка их эффективности в литературе не представлена.

Очевидно, что результат суммаризации текста зависит от конкретной постановки задачи. Согласно обзору [11], при суммаризации форумов преобладают такие задачи, как сентимент-анализ, выделение фактографической информации, анализ активности пользователей, в то же время задача выделения профессионально значимой информации не представлена даже в плане постановки. Таким образом, в настоящей работе рассматривается задача автоматической суммаризации веб-форумов с целью выделения постов, содержащих ПЗИ. Экспериментально исследуются зависимости эффективности суммаризации от контекста запроса, от применяемого метода МО, от процедуры извлечения ключевых слов.

Методика эксперимента

Отбор форумов. Для исследований были отобраны технические веб-форумы различной тематики, а на них – треды, темы которых соответствовали искомой ПЗИ (табл. 1). Длина каждого выделенного треда – не менее 400 постов.

№	Название форума	URL	Тема треда / Запрос
1	iXBT (Hardware forum)	http://forum.ixbt.com/	Choosing of ADSL modem / Как выбрать ADSL-роутер
2	Fashion, style, health	http://mail.figgery.com/	Diets for overweight people / Как сбросить лишний вес
3	Кинопоиск	http://forum.kinopoisk.ru/	«Sex at the city» сериал / Насколько хорош сериал «Секс в большом городе (Sex and the City)» и почему
4	Дом и стройка	https://www.forumhouse.ru/	Дом из бруса 200×200 с пристроем / Как построить дом из бруса 200×200 с пристроем
5	Velomania.ru	http://forum.velomania.ru/	Почему поршни возвращаются в калипер? / Почему поршни возвращаются в калипер?
6	GuitarPlayer. Ru -Форумы для гитаристов	http://forum.guitarplayer.ru/	Настройка гитары: все вопросы сюда / Как настроить гитару?
7	Вечерние платья	http://club.osinka.ru/topic-1978	Свадебное платье / Как построить выкройку корсета?
8	Sewing the wedding	http://www.thesewingforum.co.uk/showthread.php?t=37284	Dress for friends wedding - tips for sewing satin / Как обрабатывать шелковое платье?

Таблица 1. Характеристики исследуемых веб-форумов

Формирование запроса и разметка форумов. Ключевым моментом в задаче суммаризация веб-форума, как и в любой задаче извлечения информации, является оценка качества решения. Сегодня общепризнано [14], что оно должно оцениваться по отношению к информационной потребности пользователей. В связи с концептуальной сложностью и многоаспектностью оценки здесь наблюдается большое разнообразие подходов и терминологии [18–20][1].

С одной стороны, предлагается рассматривать отдельные аспекты информационной потребности. Так, автор [19] выделил шесть возможных уровней оценки информационных систем, причем три нижних описывают программно-технологические аспекты (такие как скорость выполнения запроса, построковое совпадение запроса и найденного документа и др.), а три верхних ориентированы на реакцию пользователей (в том числе наличие обратной связи, учет контекста, социальное и когнитивное соответствие запроса и документа и др.). Автор [20] использует следующие измерения: (1) характеристики пользователей (пол, возраст и пр.); (2) параметры интерактивности (число посланных запросов, просмотренных документов и пр.); (3) количественные характеристики результатов запроса (точность, полнота, *NDCG* и пр.); (4) качественные характеристики удовлетворенности пользователей (выражаемые через экспертные оценки).

С другой стороны, существует целый ряд международных проектов [14, 18, 21, 22] для оценки систем извлечения информации по типизированным информационным потребностям. Каждый проект содержит аннотированную коллекцию документов (в основном новостного характера), разделенную по группам информационных потребностей (трекам). Результаты работы систем оцениваются экспертами в лабораторных условиях по жестко установленным и контекстно-ограниченным формулировкам запросов, что ограничивает применимость такого подхода в практических задачах. К тому же треки, соответствующие извлечению информации из Интернет-форумах, в этих проектах отсутствуют.

Анализ указанных подходов показывает, что традиционные для систем извлечения информации метрики качества (такие как *F*-мера, *NDCG* и пр.) легко соотносятся с нижними уровнями классификаций [19] и [20], в то время как возможности их использования для верхних уровней связаны с адекватной формулировкой информационного запроса, предлагаемой эксперту. В связи с этим для оценки эффективности извлечения ПЗИ из веб-форумов мы формулируем информационные потребности в виде проблемно-ориентированных запросов и используем разные контексты оценки для их шкалирования (табл. 2). С одной стороны, такой подход соответствует структуре шкалирования запросов, принятой в TREC [18], с другой стороны, их содержание, как показывает практика, покрывает реальные потребности специалистов при поиске ПЗИ. Отметим, что хотя выбранные контексты являются достаточно показательными, рассмотренный подход может быть распространен и на другие контексты поиска ПЗИ на

[1] Отметим, что терминология еще не устоялась; так, в англоязычной литературе используются такие термины, как attitudes, efficiency, effectiveness, informativeness, performance, pertinence, precision, recall, relevance, satisfaction, suggestions, utility, usefulness, usability, users search success, etc.; в русскоязычной – информативность, пертинентность, полнота, релевантность, точность, эффективность, юзабилити и др.; при этом общепринятые трактовки терминов и взаимно однозначное соответствие между ними отсутствуют.

форуме. Соответствующие целевые переменные обозначаем терминами «релевантность» (relevance) и «информативность» (informativeness).

В табл. 2 представлены выбранные контексты запросов и шкалы для их оценок. Очевидно, что бинарная оценка качества извлечения ПЗИ была бы слишком грубой. Для экспертных оценок информативности и релевантности мы используем шестизначные шкалы, построенные аналогично методике [18]. Это позволяет с достаточной для практики точностью рассматривать измеряемые величины либо как категориальные, либо как непрерывные на интервале [0, 5], в зависимости от постановки задачи МО – классификация или регрессия соответственно.

Целевая переменная	Контекст оценки	Оценка	Критерий
информативность, *informativeness*	Отобрать посты, содержащие объективную, интересную и профессионально значимую информацию по теме запроса	0	Пост не содержит полезной информации
		1	Пост содержит некоторую информацию, но большая часть ее бесполезна
		2	Пост содержит определенную долю полезной информации
		3	Пост содержит значительную долю полезной информации, но отсутствуют объяснения и аргументация
		4	Пост содержит полезную информацию, но объяснения и аргументация неполны
		5	Пост содержит большой объем полезной информации с объяснениями и аргументацией
релевантность, *relevance*	Отобрать посты, содержащие информацию, относящуюся к теме запроса	0	Пост не содержит информации по теме запроса
		1	Пост слабо соответствует теме запроса
		2	В основном информация не соответствует теме запроса, но есть фрагменты, соответствующие теме запроса
		3	В основном информация соответствует теме запроса, но есть фрагменты, не соответствующие теме запроса
		4	Пост соответствует теме запроса, но есть дополнительная информация
		5	Пост полностью соответствует теме запроса

Таблица 2. Запросы и шкалы для их оценки

Отбор признаков для МО. В литературе [23–27] для извлечения информации из веб-форумов предложен спектр признаков. Опираясь на наши предыдущие работы [8–10], мы осуществили проблемно-ориентированный отбор признаков (табл. 3), которые можно условно разбить на четыре группы: (1) положение автора поста среди других пользователей (его позиция в социальном графе); (2) положение поста в треде; (3) текстовые характеристики поста; (4) эмоциональность поста. Алгоритмы оценки всех признаков, кроме последнего, подробно описаны в нашей работе [10]. В исследовании мы использовали экспертные оценки эмоциональности постов по пятибалльной шкале; в практических приложениях целесообразно определять этот признак алгоритмически (соответствующие методики широко представлены в литературе [28] и в программном обеспечении[1]).

Чтобы учесть возможную корреляцию между эмоциональной окраской и уровнем ПЗИ в посте, мы определяли веса ребер социального графа через значения эмоциональной компоненты соответствующего поста и рассчитывали признаки группы (1) двумя способами – по взвешенному (sentiment graph) и по невзвешенному (not-sentiment graph) графам.

Выбор метрик качества. Многозначная оценка эффективности извлечения ПЗИ (см. табл. 2) сужает диапазон возможных метрик качества суммаризации. Широко используемые метрики, такие как recall/precision, *F*-мера и др., и их многозначные расширения, такие как микро- и макро-*F*-меры, адекватны только для задачи классификации [14], а различные коэффициенты детерминации – для задач регрессии [13]. Следуя рекомендациям [14, 18], в качестве компромиссного варианта, применимого для обоих типов задачи МО, мы используем кумулятивную метрику [29] *NDCG* (Normalized Discounted Cumulative Gain):

[1] http://text-processing.com/docs/sentiment.html

$$NDCG = \frac{1}{CG_{\max}^{N}} \sum_{i=1}^{N} \frac{Utility_i}{\log_2(i)}$$

– метрика качества суммаризации форума, основанная на сравнении позиции текущего поста с его позицией при идеальной сортировке. Здесь $Utility_i$ – соответствующим образом определенная полезность i-го поста в отсортированной выборке; N – количество извлеченных постов; $CG_{\max}^{N}$ – сумма N наибольших значений полезности из всей выборки; суммирование производится по дисконтированным (деленным на $\log_2(i)$) значениям полезности.

Формирование обучающих выборок проводилось методом бутстреппинга [12] с числом итераций k=200, после обучения моделей результаты по всем итерациям и всем форумам усреднялись с расчетом дисперсии:

$$StD_{NDCG} = \sqrt{\frac{\sum_{i=1}^{k}\left(NDCG_i - \overline{NDCG}\right)^2}{k}},$$

где k – число итераций бутстрепа; $\overline{NDCG}$ – среднее арифметическое k значений $NDCG$; $NDCG_i$ – значение $NDCG$, полученное на i-й итерации бутстрепа.

Группа	Признак и обозначение	Содержательная трактовка
(1) позиция автора поста в социальном графе	центральность по промежуточности для невзвешенного графа (Betweeness)	значимость автора в социальной сети форума
	центральность по промежуточности для взвешенного графа (BetweenessSent)	значимость автора в социальной сети форума с учетом эмоциональных оценок
	полустепень захода вершины для невзвешенного графа (Author_inDegree)	сколько раз автор поста был процитирован
	полустепень захода вершины для взвешенного графа (Author_ inDegreeSent)	с какой эмоциональной оценкой автор поста был процитирован
	полустепень исхода вершины для невзвешенного графа (Author_outDegree)	сколько раз автор поста цитировал других авторов
	полустепень исхода вершины для взвешенного графа (Author_outDegree)	с какой эмоциональной оценкой автор поста цитировал других
	число тредов форума, в которых участвует автор поста (Number)	активность автора в социальной сети форума
(2) положение поста в треде	позиция поста относительно начала треда (Position_in_thread)	вероятность содержания офф-топика в посте
	число цитирований поста	значимость поста в треде
(3) текстовые характеристики поста	длина поста (Length)	уровень аргументации и объяснений
	наличие внешних связей (Links)	аргументация внешними источниками, изображениями и пр.
	число ключевых слов (Query_keyword_count)	соответствие теме
	число наиболее частотных ключевых слов (Most_used_topic_keyword_count)	соответствие теме
(4) эмоциональность поста	оценка по шкале –2...+2	пост содержит явно выраженные позитивные ... негативные эмоции

Таблица 3. Признаки, используемые для машинного обучения

Отбор методов и моделей МО. К настоящему времени разработано большое количество методов МО и реализующих их алгоритмов. В академической литературе [12, 13, 30] используются различные принципы их классификации; в настоящей работе мы используем терминологию [12]. Кроме того, постоянно появляются их рейтинги, составляемые пользователями и (или) разработчиками прикладных программных пакетов[1,2].

Как показано выше, результат извлечения ПЗИ определяется контекстом запроса и типом метрики оценки, которая, в свою очередь, связана с постановкой задачи МО. Исходя из этого, для отбора методов

[1] Топ-10 data mining-алгоритмов простым языком // https://habrahabr.ru/company/itinvest/blog/262155/
[2] Выбор алгоритмов машинного обучения Microsoft Azure // https://azure.microsoft.com/ru-ru/documentation/articles/machine-learning-algorithm-choice/

и моделей МО в нашем случае мы использовали два классифицирующих признака: тип задачи МО (классификация / регрессия) и целевая переменная, выбранная для характеристики эффективности извлечения ПЗИ (информативность / релевантность).

Результаты и обсуждение

Зависимость эффективности суммаризации форумов от типа задачи МО. На рис. 1 и 2 представлены результаты применения доступных в пакете Weka реализаций алгоритмов МО к форумам №№ 5 и 6 соответственно. Эксперименты проводились на алгоритмах МО, представленных в пакете Weka v. 3.6.13[1]. Использована разметка форумов в шкале информативности.

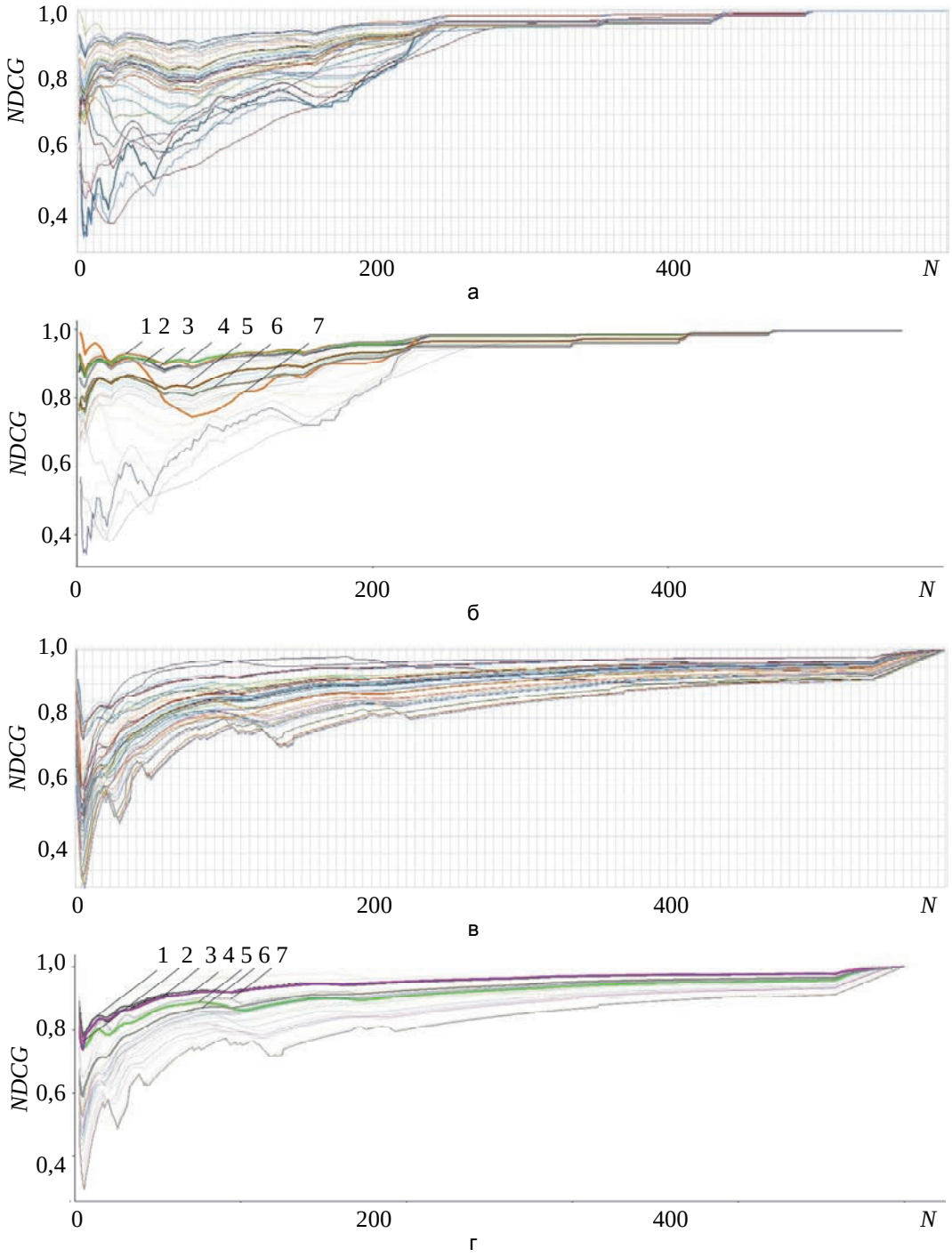

Рис. 1. Зависимости метрики *NDCG* от количества извлеченных постов *N* для различных алгоритмов машинного обучения; тип задачи – классификация, целевая переменная – информативность: форум № 5 (а)–(б);– форум № 6 (в)–(г); 57 алгоритмов (а)–(в); 7 алгоритмов (б)–(г)

[1] Weka v. 3.6.13 // http://www.cs.waikato.ac.nz/ml/weka/

Метод МО	Модель алгоритма, тип алгоритма, номер на рис. 1, 2
Метрический	1. Модель ближайших соседей, IBk. № 3 на рис. 1, б, № 3 на рис. 1, г
	2. Модель ближайших соседей с обобщением, NNge. № 4 на рис. 1, б, № 1 на рис. 1, г
Алгоритмическая композиция	3. Линейная композиция классификаторов, RandomCommittee. № 1 на рис. 1, б, № 2 на рис. 1, г
Бустинг	4. Бустинг, LogitBoost. № 6 на рис. 1б, № 5 на рис. 1, г
Логическая классификация	5. Решающий список на основе дерева C4.5, PART. № 5 на рис. 1, б, № 6 на рис. 1, г
	6. Дерево решений, RandomTree. №2 на рис. 1, б, №4 на рис. 1, г
Метод опорных векторов	7. Радиально-базисный метод опорных векторов, LibSVM. № 7 на рис. 1, б, № 7 на рис. 1, г

Таблица 4. Алгоритмы, показавшие лучшие результаты при анализе форумов №№ 5, 6 (табл. 1)

Классификационная или регрессионная модель алгоритма	Тип алгоритма, оптимизирующие параметры	
	классификации	регрессии
Линейная регрессия в режиме регрессии / Логистическая регрессия в режиме классификации	Logistic	LinearRegression
	By default	By default
Дерево решений (C4.5 и M5)	J48	M5P
	Use reduced error pruning: true	Build regression tree/rule rather than a model tree/rule: true
Радиально-базисный метод опорных векторов в режиме классификации / SVM-регрессия в режиме регрессии	LibSVM	LibSVM
	By default	By default
Метод ближайших соседей	IBk	IBk
	Neighbors number: 5 Weight neighbors: by the inverse of their distance Neighbor's number selection: hold-one-out evaluation Minimization parameter: mean squared error	Neighbors number: 5 Weight neighbors: by the inverse of their distance Neighbor's number selection: hold-one-out evaluation Minimization parameter: mean squared error
Нейронная сеть	MultilayerPerceptron	MultilayerPerceptron
	Learning Rate for the backpropagation algorithm: 0.001 Momentum Rate for the backpropagation algorithm: 0.001 Number of epochs to train through: 5000 Percentage size of validation set: 20	Learning Rate for the backpropagation algorithm: 0.001 Momentum Rate for the backpropagation algorithm: 0.001 Number of epochs to train through: 5000 Percentage size of validation set: 20
Наивный Байес с использованием ядер в режиме классификации / Гауссовская ядерная модель в режиме регрессии	NaiveBayes	GaussianProcesses
	Use kernel density estimator: true	By default

Таблица 5. Сопоставленные алгоритмы классификации и регрессии

В первых экспериментах к каждому форуму по отдельности применялись все (57 шт.) доступные в пакете Weka алгоритмы МО. Однако, как видно из рис. 1, а, в, получаемые таким образом результаты, несмотря на подробность и преимущества визуализации, трудно интерпретируются при сравнительном анализе. Кроме того, была выявлена достаточно слабая согласованность между ранжировками алгоритмов, получаемыми по отдельным форумам. Например, коэффициент ранговой корреляции Кендалла для

результатов обучения по форуму № 5 (рис. 1, а) и по форуму № 6 (рис. 1, в) составил $\tau_{57} = 0{,}55$. Алгоритмы (7 шт.), показавшие наилучшую эффективность и в то же время достаточную согласованность при обучении по этим форумам ($\tau_7 = 0{,}7$), представлены на рис. 1, б, г, и в табл. 4.

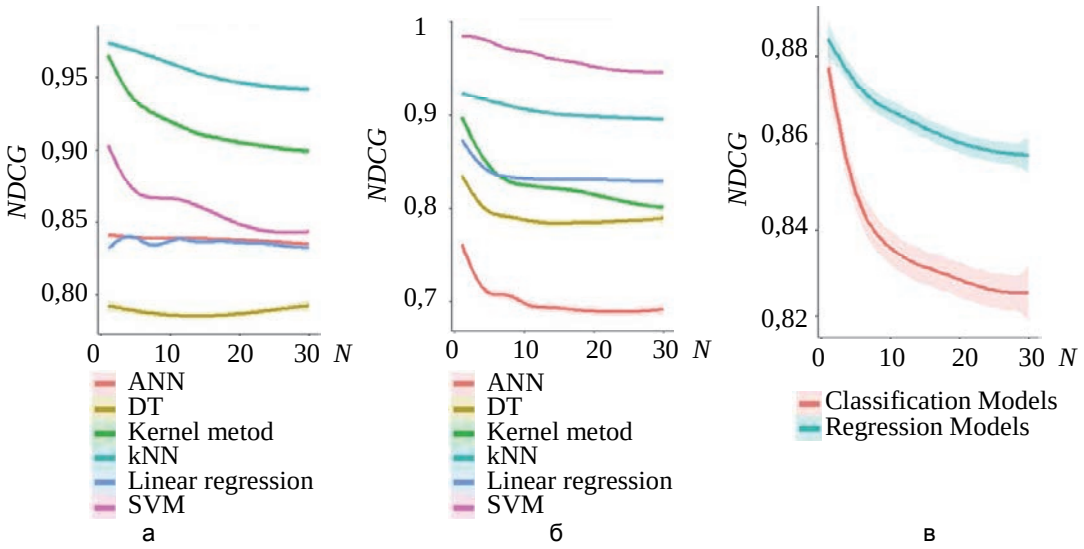

Рис. 2. Зависимости метрики *NDCG* и ее дисперсии StD_{NDCG} (показана светлым тоном) от количества извлеченных постов *N* для различных типов задачи машинного обучения; алгоритмы – по табл. 5; форумы – №№ 5–8 (табл. 1); целевая переменная – информативность: тип задачи – регрессия, усреднение по форумам для каждого алгоритма (а); тип задачи – классификация, усреднение по форумам для каждого алгоритма (б); типы задачи – классификация и регрессия, усреднение по всем алгоритмам (в)

Для дальнейшего исследования использованы шесть наиболее рейтинговых алгоритмов классификации, к каждому из которых отобрана аналогичная регрессионная модель МО (табл. 5). В табл. 5 указаны значения параметров настройки, полученные при оптимизации каждого алгоритма; для остальных параметров использованы значения по умолчанию. Результаты экспериментов, сгруппированные по целевой переменной, представлены на рис. 2, а–в (целевая переменная – информативность), и на рис. 3, а–в (целевая переменная – релевантность). Диапазон значений *NDCG* (*N*=0–30) на рис. 2, 3 соответствует начальным, наиболее информативным фрагментам рис. 1.

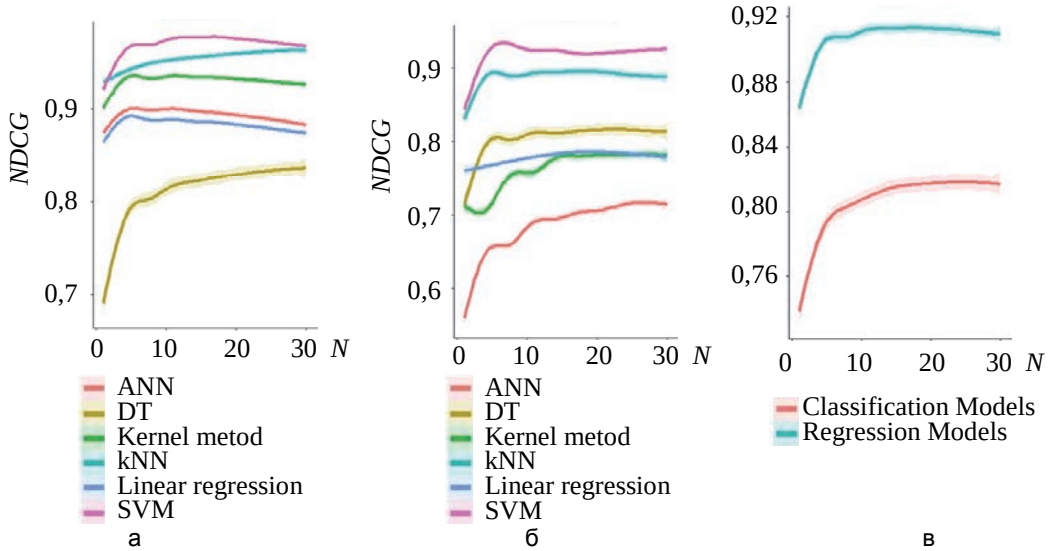

Рис. 3. Зависимости метрики *NDCG* и ее дисперсии StD_{NDCG} (показана светлым тоном) от количества извлеченных постов *N* для различных типов задачи машинного обучения; алгоритмы – по табл. 5; форумы – №№ 1–4 (табл. 1); целевая переменная – релевантность:– тип задачи – регрессия, усреднение по форумам для каждого алгоритма (а); тип задачи – классификация, усреднение по форумам для каждого алгоритма (б); модели – классификация и регрессия, усреднение по всем алгоритмам (в)

Сопоставление рис. 2, в, и рис. 3, в, показывает, что, независимо от целевой переменной, эффективность суммаризации в режиме регрессии выше, чем в режиме классификации. Этот результат был подтвержден экспериментами авторов на других датасетах; он теоретически обоснован для алгоритма

SVM в работе [31]. Подчеркнем, что он получен для шестиуровневой шкалы оценки целевых переменных (табл. 2), адекватной при извлечении ПЗИ; при оценке релевантности в двухуровневой шкале, которая характерна для некоторых задач извлечения информации, режим классификации может оказаться более эффективным [14]. Сопоставление рис. 2, а, и рис. 3, а, не позволяет выделить наилучший алгоритм для извлечения ПЗИ, однако среди шести исследованных алгоритмов наблюдается сравнительно высокая согласованность (коэффициент ранговой корреляции Кендалла составил $\tau_6 = 0{,}73$).

Зависимость эффективности суммаризации форумов от контекста запроса. Величина коэффициента корреляции между результатами экспертной разметки информативности и релевантности, усредненная по всем форумам (табл. 1), составила K=0,36. Тем самым подтверждается, что результаты суммаризации форумов, выполняемой экспертами вручную, существенно зависят от контекста запроса. Для более детальной оценки этих различий построены усредненные по всем форумам распределения оценок информативности и релевантности (рис. 4, а, б, соответственно). Как видно из рисунков, оба распределения отличаются от нормального. Распределение информативности, получаемое по оценкам экспертов, имеет максимум в области *informativeness* = 3, и при этом доля крайних значений (как очень информативных, так и неинформативных постов) весьма мала, в то же время распределение релевантности скошено в сторону *relevance* = 5. Такой скос можно объяснить процедурой отбора постов из заведомо более релевантных тредов, чем весь форум целиком.

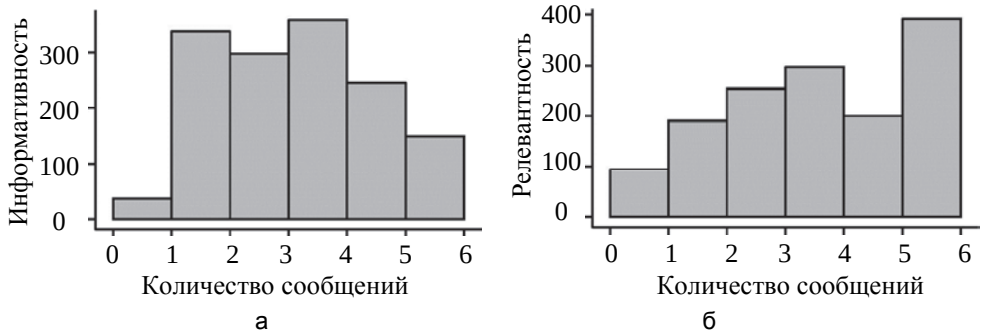

Рис. 4. Распределения экспертных оценок эффективности извлечения профессионально значимой информации в зависимости от контекста запроса: целевая переменная – информативность (а); целевая переменная – релевантность (б)

Для анализа влияния контекста запроса на эффективность машинного обучения использованы следующие модели:

- множественная линейная регрессия (МЛР) – допускает содержательную математическую интерпретацию, хорошо описывает линейные зависимости; в связи с достаточным объемом выборки и слабой корреляцией между признаками используется модель без регуляризации;
- стохастический градиентный бустинг (СГБ, англ. Stochastic Gradient Boosting) – допускает эвристическую интерпретацию, хорошо описывает нелинейные зависимости;
- латентное размещение Дирихле (LDA, англ. Latent Dirichlet Allocation) – робастная, неинтерпретируемая. Модель рассматривает текст треда как коллекцию документов (постов), причем каждый документ с определенной вероятностью относится к нескольким неявно выраженным (латентным) темам, а каждая тема связывается с определенным набором слов из треда. Сравнивая наборы слов в сформированных темах и в постах, размеченных экспертами, можно выделить посты, соответствующие конкретному контексту запроса.

Параметры настройки алгоритмов приведены в табл. 6. При бутстреппинге использовалось деление выборки на обучающую и контрольную в соотношении 70%:30%.

Модель	Средство реализации	Параметры настройки
МЛР	функция lm пакета stats библиотеки языка R v3.0.2	–
СГБ	пакет gbm v.2.1	number of trees = 2000, shrinkage factor = 0,001, number of splits for each tree = 3
LDA	Библиотека Mallet для языка R v1.0	100 iterations and 3 topics

Таблица 6. Параметры настройки алгоритмов при сравнении контекстов запроса

На рис. 5 представлены оценки качества суммаризации, выполняемой отдельными алгоритмами, для разных характеристик эффективности извлечения ПЗИ, которые, в свою очередь, связаны с контекстами запроса. Сопоставляя рис. 2, а, и рис. 5, а, с рис. 3, а, и рис. 5, б, можно видеть, что при оценке релевантности линейные методы проявляют себя лучше, чем нелинейные, а при оценке информативности зависимость обратная и выражена слабее. При этом, как показали дополнительные исследования, харак-

тер нелинейности зависит от конкретных признаков: так, наблюдается сильная корреляционная зависимость между длиной текста поста (Length, табл. 3) и информативностью поста.

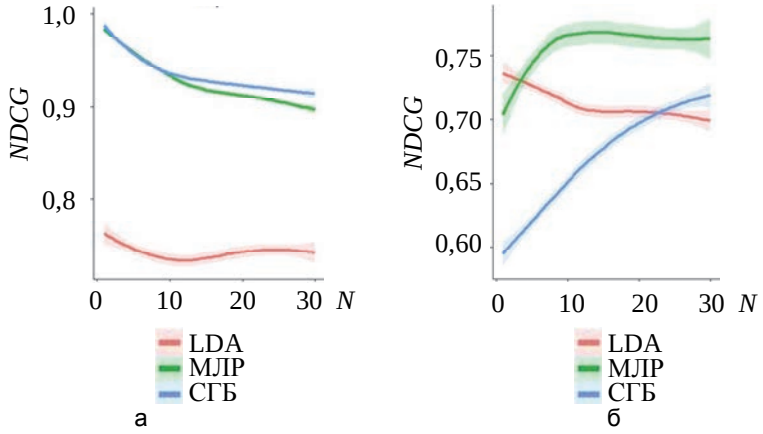

Рис. 5. Зависимости метрики *NDCG* и ее дисперсии StD_{NDCG} (показана светлым тоном) от количества извлеченных постов *N* для различных контекстов запроса; алгоритмы – согласно табл. 6; доверительный интервал для всех кривых 99,5%: целевая переменная – информативность (а); целевая переменная – релевантность (б)

Зависимость эффективности суммаризации форумов от процедуры извлечения ключевых слов. Естественно предположить, что при суммаризации форума как текстового документа признаки, связанные с отбором ключевых слов, будут играть доминирующую роль. В пользу этого предположения свидетельствует, в частности, доминирование модели «bag-of-words» в большинстве задач интеллектуальной обработки текстов [14]. Однако наши предыдущие работы [8–10] не подтверждают этот тезис: для различных методов и алгоритмов МО, примененных при суммаризации форумов, топ-лист признаков оказывается различным, причем текстовые характеристики поста не всегда занимают первые места в списке. В качестве примера в табл. 7 приведены фрагменты ранжировки признаков для лучших алгоритмов из рис. 4. Отбор производился с уровнем значимости 0,001.

Алгоритм	МЛР	СГБ
Целевая переменная	релевантность	информативность
Признаки в порядке убывания значимости	Query_keyword_count Most_used_topic_keyword_count Author_inDegree Author_inDegreeSent	Length Author_outDegree Author_outDegreeSent Position_in_thread

Табл. 7. Фрагменты ранжировки признаков (позиции 1–4) для лучших алгоритмов из рис. 4

Данные табл. 7 показывают, что при оценке релевантности ведущую роль играют признаки, связанные с ключевыми словами, а при оценке информативности на первый план выступают характеристики текста поста в целом (длина поста и его положение в треде), а также признаки, связанные со структурой треда как текста и как социального графа. Отметим, что аналогичный результат был получен в [10].

Для более детального исследования этого вопроса мы сравнили результаты суммаризации форумов при использовании различных способов выделения ключевых слов, в том числе:

- most used keywords. Модель рассматривает посты форума как общий набор слов и вычисляет среди него те, частота повторений которых максимальна;
- hclust. Модель рассматривает текст треда как матрицу «термин-документ». В начале работы алгоритма каждое слово относится к отдельному кластеру. Далее на каждой итерации выбирается два наиболее близких кластера и стягиваются в один, т.е. образуется иерархическая классификация слов. На заключительном этапе, в зависимости от параметров настройки моделей, можно однозначно определить, какому кластеру принадлежат слова, и выбрать наиболее встречаемые в каждом;
- Kmeans – Модель рассматривает текст треда как матрицу «термин-документ» и относит каждый термин к тому подмножеству (кластеру), расстояние до центра которого минимально;
- expert – экспертное выделение ключевых слов. Выделяются слова, несущие в себе максимальную смысловую нагрузку относительно темы треда на форуме;
- LDA– робастная, неинтерпретируемая.

Параметры настройки алгоритмов приведены в табл. 8. Во всех моделях при обработке сообщений из текста были предварительно удалены стоп-слова.

Модель	Средство реализации	Параметры настройки
most used keywords	Самостоятельно реализованная функция	–
hclust	Функция hclust() пакета stats для языка R	distance="binary", cluster.count = 3, keyword.count.per.cluster = 3
kmeans	Функция kmeans() пакета stats для языка R	cluster.count = 3, keyword.count.per.cluster = 3

Таблица 8. Параметры настройки алгоритмов при сравнении способов отбора ключевых слов

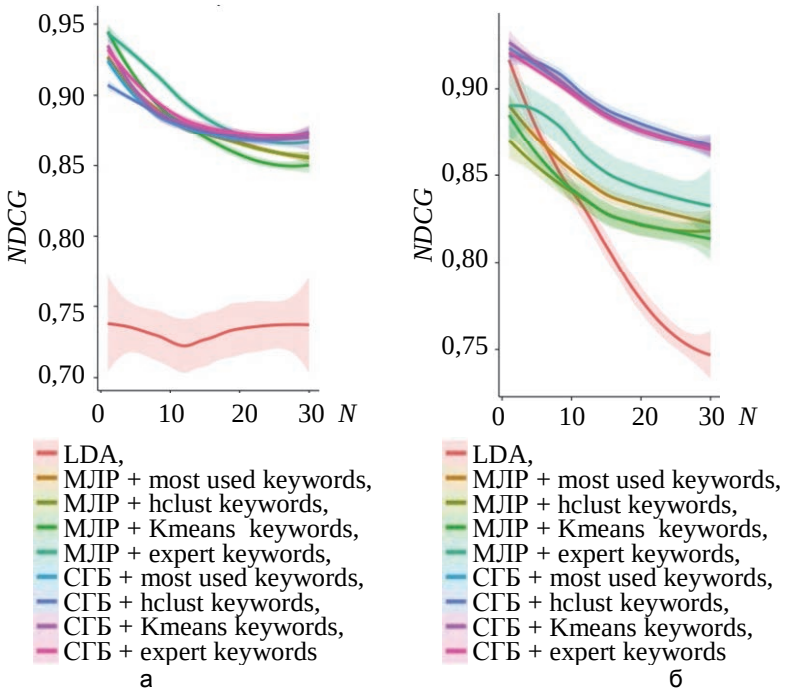

Рис. 6. Зависимости метрики *NDCG* и ее дисперсии StD_{NDCG} (показана светлым тоном) от количества извлеченных постов *N* для различных методов извлечения ключевых слов; алгоритмы – согласно табл. 6: целевая переменная – информативность, усреднение по 3 форумам для каждого алгоритма (а); целевая переменная – релевантность, усреднение по форумам для каждого алгоритма (б)

Рис. 6 подтверждает вывод, сделанный по табл. 7: эффективность извлечения информативных постов (рис. 6, а) практически не зависит от способа задания ключевых слов, в то время как для извлечения релевантных постов (рис. 6, б) такая зависимость существенна. При этом способы most used keywords и expert keywords оказались почти идентичными по эффективности (рис. 6, б), что может упростить автоматизацию выделения ПЗИ в реальных приложениях. Кроме того, сопоставляя рис. 5 и 6, можно заметить, что при выделении релевантных постов модель LDA занимает промежуточное положение между линейными и нелинейными методами (рис. 5, б, рис. 6, б), а при выделении информативных постов модель LDA значительно уступает им обоим (рис. 5, а, рис. 6, а).

Полученные результаты можно интерпретировать следующим образом. Релевантность постов форума в решающей степени определяется их словарным составом, информативность постов форума в значительно большей мере, чем релевантность, связана с семантикой форума в целом и выражается через характеристики связности текстов постов как лингвистической структуры, а также связности социального графа форума. Поэтому для выделения релевантных постов вполне адекватными могут оказаться методы, основанные на модели «bag-of-words», в том числе классические методы текстового поиска по ключевым словам или тематического моделирования (типа LDA). В то же время для извлечения информативных постов целесообразно использовать специализированные алгоритмы, основанные на рассмотренных в работе принципах.

Очевидный интерес для реальных систем представляют обобщенные алгоритмы извлечения информации для различных типов запроса [32]. Результаты настоящей работы позволяют варьировать не только тип, но и контекст запроса, что необходимо при извлечении ПЗИ.

Заключение

Рассмотрена задача автоматической суммаризации веб-форумов с целью выделения постов, содержащих профессионально значимую информацию. Показано, что в оценке эффективности извлечения

профессионально значимой информации важную роль играет контекст запроса. Отобраны характерные для извлечения профессионально значимой информации контексты запросов, отражающие различные трактовки информационной потребности пользователей. Для обозначения оцениваемых характеристик качества извлечения профессионально значимой информации использованы термины «релевантность» (relevance) и «информативность» (informativeness). Построены шкалы для их оценок, соответствующие общемировым подходам.

Для исследований отобраны шесть веб-форумов, тематикой которых являются аспекты технологий различных предметных областей. Разметка форумов проводилась экспертным путем. С использованием различных методов машинного обучения построены модели, отражающие функциональную связь между оцениваемыми характеристиками качества извлечения профессионально значимой информации и признаками постов. Для оценки качества моделей использованы кумулятивная метрика *NDCG* и ее дисперсия.

Экспериментально подтверждено, что результаты суммаризации форумов, выполняемой экспертами вручную, существенно зависят от контекста запроса. Показано, что, независимо от целевой переменной, эффективность извлечения профессионально значимой информации в режиме регрессии выше, чем в режиме классификации.

Показано, что в общей оценке эффективности извлечения профессионально значимой информации релевантность достаточно хорошо описывается линейной комбинацией признаков, а для оценки информативности уже требуется их нелинейная комбинация. При этом при оценке релевантности ведущую роль играют признаки, связанные с ключевыми словами, а при оценке информативности на первый план выступают характеристики текста поста в целом, а также признаки, связанные со структурой треда как текста и как социального графа.

Показано, что эффективность извлечения информативных постов слабо зависит от способа задания ключевых слов, в то время как для извлечения релевантных постов такая зависимость существенна. Выявлен способ выделения ключевых слов, наиболее эффективный для реальных приложений. Показано, что при выделении релевантных постов линейные методы выигрывают в эффективности по сравнению с нелинейными, а модель LDA занимает промежуточное положение; в то же время при выделении информативных постов линейные и нелинейные методы идентичны по эффективности, а модель LDA значительно уступает им обоим.

Предложена содержательная модель, позволяющая объяснить полученные результаты.

Полученные результаты могут служить основой для построения и новых и адекватного применения существующих алгоритмов суммаризации веб-форумов, что позволит существенно сократить временные и ресурсные затраты пользователя на получение и изучение информации максимально свежей профессионально значимой информации.

References

1. Vasiliev V.N., Lisitsyna L.S. Planning and estimation of expected competences learning outcomes for FSES HPE. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2013, no. 2 (84), pp. 142–148. (In Russian)
2. Vasiliev V.N., Lisitsyna L.S., Shehonin A.A. Conceptual model for the extraction of learning outcomes from the excessive education content *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2010, no. 4, pp. 104–108. (In Russian)
3. Lisitsyna L.S. *Methodology of Designing Modular Competence-Oriented Education Programs*. St. Petersburg, SPbSU ITMO, 2009, 50 p. (In Russian)
4. Druzhinin V.N. Psychology. 2nd ed. St. Petersburg, Piter Publ., 2009, 656 p.
5. *Kontseptsiya i metodika razrabotki kontrol'no-otsenochnykh sredstv*. Available at: http://www.firo.ru/wp-content/uploads/2012/12/Concetion.doc (accessed 29.04.2016).
6. Stolyarenko A.M. *Psychology and Pedagogics*. 3rd ed. Moscow, Yuniti-Dana, 2010, 544 p. (In Russian)
7. Gusarova N.F., Kovalenko M.N., Mayatin A.V., Petrov V.A., Shilov I.V. Using a hierarchically organized text online forum as a means to support the scientific and technical design. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2005, no. 20, pp. 243–247. (In Russian)
8. Grozin V.A., Dobrenko N.V., Gusarova N. F., Tao N. The application of machine learning methods for analysis of text forums for creating learning objects. *Proc. Int. Conf. on Computational Linguistics and Intellectual Technologies*. Moscow, 2015, vol. 1, no. 14, pp. 202–213.
9. Grozin V.A., Gusarova N.F., Dobrenko N.V. Feature selection for language-independent text forum summarization. *Proc. 6th Int. Conf. on Knowledge Engineering and Semantic Web, KESW - 2015*. Moscow, 2015, vol. 518, pp. 63–71. doi: 10.1007/978-3-319-24543-0_5
10. Buraya K.I., Grozin V.A., Gusarova N.F, Dobrenko N.V. Machine learning methods for extracting of professionally significant information from web forums. *Distantsionnoe i Virtual'noe Obrazovanie*, 2015, no. 12, pp. 46–63.

11. Almahy I., Salim N. Web discussion summarization: study review. *Proc. 1st Int. Conf. on Advanced Data and Information Engineering, DaEng-2013*. Kuala Lumpur, Malaysia, 2013, pp. 649–656. doi: 10.1007/978-981-4585-18-7_73
12. Vorontsov K.V. *Machine Learning (Lectures)*. Available at: http://www.machinelearning.ru/wiki/index.php?title= Машинное обучение (курс лекций, К.В.Воронцов) (accessed 04.2016).
13. Bishop C.M. *Pattern Recognition and Machine Learning*. Springer, 2006, 738 p.
14. Manning C.D., Raghavan P., Schutze H. *Introduction to Informational Retrieval*. Cambridge University Press, 2008, 504p.
15. Beliga S., Mesrovic A., Martinic-Ipsic S. An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, 2015, vol. 39, no. 1, pp. 1–20.
16. Zhao H., Zeng Q. Micro-blog keyword extraction method based on graph model and semantic space. *Journal of Multimedia*, 2013, vol. 8, no. 5, pp. 611–617. doi: 10.4304/jmm.8.5.611-617
17. Sondhi P., Gupta M., Zhai C.X., Hockenmaier J. Shallow information extraction from medical forum data. *Proc. 23rd Int. Conf. on Computational Linguistics, COLING '10*. Beijing, China, 2010, pp. 1158–1166.
18. Elbedweihy K.M., Wrigley S.N., Clough P., Ciravegna F. An overview of semantic search evaluation initiatives. *Journal of Web Semantics*, 2015, vol. 30, pp. 82–105. doi: 10.1016/j.websem.2014.10.001
19. Saracevic T. Evaluation of evaluation in information retrieval. *SIGIR Forum*, 1995, pp. 137–146.
20. Kelly D. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends Information Retrieval*, 2009, vol. 3, no. 1–2, pp. 1–1224. doi: 10.1561/1500000012
21. Nenkova A., McKeown K. A survey of text summarization techniques. *Mining Text Data*, 2012, pp. 43–76. doi: 10.1007/978-1-4614-3223-4_3
22. Harman D. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 2011.
23. Biyani P., Bhati S., Caragea C., Mitra P. Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems*, 2014, vol. 69, no. 1, pp. 170–178. doi: 10.1016/j.knosys.2014.04.048
24. Smine B., Faiz R., Desclés J-P. Relevant learning objects extraction based on semantic annotation. *Interanational Journal of Metadata, Semantics and Ontologies*, 2013, vol. 8, no. 1, pp. 13–27. doi: 10.1504/IJMSO.2013.054187
25. Nettleton D.F. Data mining of social networks represented as graphs. *Computer Science Review*, 2013, vol. 7, no. 1, pp. 1–34. doi: 10.1016/j.cosrev.2012.12.001
26. Romero C., Lopez M.-I., Luna J.-M., Ventura S. Predicting students' final performance from participation in on-line discussion forums. *Computers and Education*, 2013, vol. 68, pp. 458–472. doi: 10.1016/j.compedu.2013.06.009
27. Wang B.-X., Liu B.-Q., Sun C.-J., Wang X.-L., Sun L. Thread segmentation based answer detection in Chinese online forums. *Acta Automatica Sinica*, 2013, vol. 39, no. 1, pp. 11–20. doi: 10.3724/SP.J.1004.2013.00011
28. Mihalcea R., Banea C., Wiebe J. Learning multilingual subjective language via cross-lingual projections. *Proc. 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, 2007, pp. 976–983.
29. Järvelin K., Kekäläinen J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 2002, vol. 20, no. 4, pp. 422–446. doi: 10.1145/582415.582418
30. Shai S.-S., Shai B.-D. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014, 409 p.
31. Herbrich R., Graepel T., Obermayer K. Large-margin thresholded ensembles for ordinal regression: theory and practice. In: *Advances in Large Margin Classifiers*. MIT Press, 2000, pp. 115–132.
32. Croft W.B. Combining approaches to information retrieval. In: *Advances in Information Retrieval*. Ed. W.B. Croft. Springer, 2000, pp. 1–36.

Бурая Ксения Игоревна – младший java-разработчик, ЗАО «Петер-Сервис», Санкт-Петербург, 191123, Российская Федерация, ks.buraya@gmail.com

Виноградов Павел Дмитриевич – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; инженер-программист, ООО ЛКМСПБГУ, Санкт-Петербург, 199034, Российская Федерация, Pavel.d.vinogradov@gmail.com

Грозин Владислав Андреевич – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; аналитик данных, Диджинетика, Москва, 125319, Российская Федерация, vlad.grozin@yandex.ru

Гусарова Наталия Федоровна – кандидат технических наук, старший научный сотрудник, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, natfed@list.ru

Добренко Наталья Викторовна – ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, graziokisa@gmail.com

Трофимов Владислав Александрович – студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; инженер-программист, Санкт-Петербургский Центр Разработок EMC, Санкт-Петербург, 199004, Российская Федерация, vladisl.trofimov@gmail.com

Kseniya I. Buraya – junior Java developer, Peter-Service, Saint Petersburg, 191123, Russian Federation, ks.buraya@gmail.com

Pavel D. Vinogradov – student, ITMO University, Saint Petersburg, 197101, Russian Federation; software engineer, Computer Modelling Laboratory of the St. Petersburg State University (CML SPBGU), Saint Petersburg, 199034, Russian Federation, Pavel.d.vinogradov@gmail.com

Vladislav A. Grozin – student, ITMO University, Saint Petersburg, 197101, Russian Federation; data analyst, Diginetica, Moscow, 125319, Russian Federation, vlad.grozin@yandex.ru

Natalia F. Gusarova – PhD, senior scientific researcher, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation, natfed@list.ru

Natalia V. Dobrenko – assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, graziokisa@gmail.com

Vladislav A. Trofimov – student, ITMO University, Saint Petersburg, 197101, Russian Federation; software engineer, EMC Corporation, Saint Petersburg, 199004, Russian Federation, vladisl.trofimov@gmail.com