УДК 004.912

А.А. Зайцева, С.В. Кулешов, С.Н. Михайлов

МЕТОД ОЦЕНКИ КАЧЕСТВА ТЕКСТОВ В ЗАДАЧАХ АНАЛИТИЧЕСКОГО МОНИТОРИНГА ИНФОРМАЦИОННЫХ РЕСУРСОВ

Зайцева А.А., Кулешов С.В., Михайлов С.Н. **Метод оценки качества текстов в задачах аналитического мониторинга информационных ресурсов.**
Аннотация. В статье предлагается метод оценки качества технических текстов, основанный на применении подхода аналитического самореферирования. Предлагается эвристический критерий качества текстов, основанный на оценке скорости уменьшения объема реферата текста.
Ключевые слова: текст, аналитический мониторинг, информационный ресурс, реферирование, глоссарий, критерий качества текстов, семантический анализ.

Zaytseva A.A., Kuleshov S.V., Mikhailov S.N. **The Method for the Text Quality Estimation in the Task of Analytical Monitoring of Information Resources.**
Abstract. The paper proposes the method of technical texts quality estimation based on analytical auto annotation approach. The heuristic criteria for text quality based on annotation volume reduction rate is proposed.
Keywords: text, analytical monitoring, information resource, auto annotation, glossary, quality criteria of texts, semantic analysis.

1. Введение. В задачах анализа текущего технологического состояния и прогнозирования тенденций развития различных отраслей жизнедеятельности общества, одним из первых этапов является подбор литературы по теме проводимого исследования. Традиционно этот этап выполняется вручную при помощи поисковых систем Интернет, чтения найденных электронных ресурсов, сохранения их копий (или ссылок), корректировки поискового запроса в соответствии со списком ключевых слов и фраз и повторения всех описанных действий до накопления требуемого объема текстового материала, необходимого для продолжения работы или принятия решения об отсутствии материала по выбранной тематике.

В случае, когда исследование длится достаточно долго или тема является основной для исследователя, интересным и крайне полезным является наблюдение динамики изменений (мониторинг) публикаций по интересующей тематике. Подобный мониторинг по набору заданных тем необходим и при организации учебного процесса [1], и в системах прогнозирования различных чрезвычайных ситуаций и техногенных катастроф социотехнического характера, в которых точность прогнозов сильно зависит от корректности и достоверности исходных материалов [2].

При проведении мониторинга требуется выполнять оценку качества текстов, поступающих на вход системы. Такими критериями

могут служить: связность изложения, количество вхождений ключевых слов в тексте, степень «научности» и т.д.

К синтаксическим особенностям оформления научно-технических текстов следует отнести синтаксическую полноту оформления высказывания, частое употребление стандартных паттернов, развернутую систему связующих элементов (союзов, союзных слов).

Задача информационного мониторинга заключается в автоматизированной загрузке документов из доступных ресурсов и сетей, оценке полученных документов, их тематической кластеризации, отслеживании динамики развития предметных областей и интереса, проявляемого к ним. Также могут решаться задачи создания корпусов текстов, построения тематических тезаурусов и онтологий.

При решении задачи мониторинга системе приходится обрабатывать большие объемы документов из сети Интернет, большая часть из которых оказывается текстами рекламно-информационного или развлекательного характера; служебными страницами, обеспечивающими функционирование web-сайтов или текстами SEO-направленности (search engine optimization — тексты, специально направленные на манипулирование поисковыми системами с целью повышения значимости отдельных интернет-ресурсов в результатах поисковых систем).

Для повышения эффективности работы системы целесообразно использовать входной контроль данных с превентивной отбраковкой текстов, заведомо не удовлетворяющих критерию полезности, а зачастую искажающих результаты мониторинга. Для решения подобных задач разработаны различные методы семантического анализа, к их числу следует отнести системы, построенные на основе методов машинного обучения по примерам [3]. Подобные системы могут контролировать ошибки, которые отражают не только грамотность, но и качество построения лингвистических конструкций автором: отсутствие субъекта высказывания, частое повторение в предложении и соседних предложениях определенного слова, нарушение падежного и других согласований и т.д. Их наличие выявляется с помощью специальных семантических правил [4]. Все вышеперечисленные признаки имеют отношение к качеству текста, но не являются всеобъемлющими критериями, позволяющими корректно оценить смысловую составляющую контента. Предлагаемый подход позволяет дополнить существующие методы оценкой контента с точки зрения отделения научных и технических текстов от текстов рекламного характера, поискового спама и т.д.

2. Реферирование. Задача реферирования текста стоит очень остро в любом большом хранилище документов, в том числе в локаль-

ных и глобальных сетях. Наличие реферата–аннотации к документу, составленного из элементов текста значительно упрощает задачу поиска документов.

Идея автоматического построения реферата к произвольному тексту основана на построении понятийного окружения (рисунок 1) терминов текста аналогично формированию визуального глоссария [5]. После статистической фильтрации семантического окружения (порог которой определяет степень сжатия текста полученного реферата) производится обратное преобразование понятийного окружения, представляющего граф, вершинами которого являются слова, в текст. При этом формируется сжатый пересказ текста, прообразом которого является исходный текст обрабатываемого документа [6].

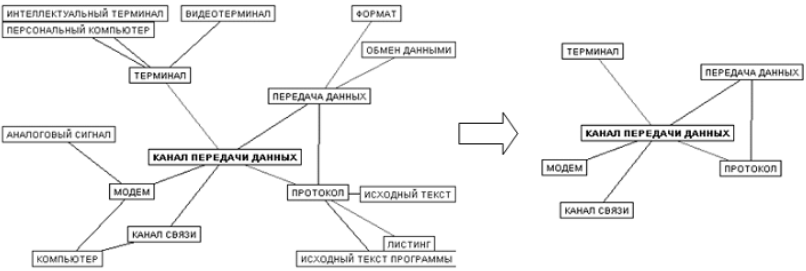

Рис. 1. Пример уменьшения понятийного окружения текста

При этом основная семантическая составляющая, проходящая через весь текст и многократно подчеркиваемая (возможно, в разных близких терминах) оказывается вынесенной в реферат, а все второстепенные описания отбрасываются [6].

Описанный метод позволяет выбрать из текста предложение (или набор предложений), наиболее полно отражающее содержимое текста, т.е. предложение (предложения), содержащее максимальное количество словесных пар (связей), имевших наибольший статистический вес при разборе. При этом порог веса является параметром компрессии получаемого реферата.

3. Оценка качества текстов. При современном состоянии информационной составляющей Интернет, имеющей явную рекламно-коммерческую и развлекательную направленность, основной ошибкой систем автоматического мониторинга и информационного поиска является включение в архив документов и поисковый индекс бесполезных, а зачастую и «вредных» ресурсов, искажающих общую информационную картину.

Усугубляет ситуацию использование «спамогенераторов» и «синонимайзеров» — программ для повышения уникальности текстов посредством замены отдельных слов или фраз в них на синонимы (аналогичные по смыслу слова или фразы). Некоторые варианты программ подобного типа [7] умеют перефразировать предложения, менять местами абзацы и т.д.

Как указано в [8], в настоящее время существует несколько подходов к созданию текстов для спам-сайтов:

– создание текстов вручную;

– копирование текстов из других источников;

– автоматическая генерация текстов;

– автоматическая модификация существующих текстов;

– использование автоматического перевода текстов на другой язык и обратно;

– различные методы уникализации текстов [9].

Для уменьшения влияния такого рода контента в задачах информационного мониторинга интернет-ресурсов выполняется оценка качества текстов, поступающих на вход системы. Такими критериями могут служить как внешние показатели качества (индексы цитирования [10], количество внешних ссылок, PageRank [11]), косвенно зависящие от самого текстового контента документа, так и собственные характеристики текста: связность изложения, «научность», количество вхождений ключевых слов в тексте и т.д. [12].

Для повышения качества отделения научно-технических текстов от текстов рекламного характера предлагается метод, основанный на оценке скорости уменьшения объема автоматически сформированного реферата текста на каждом шаге реферирования (при последовательном увеличении порога ε).

Пусть $s \in \mathrm{T}$, где s — предложения, являющиеся элементами множества предложений текста T.

В этом случае рефератом текста назовем множество F, если $\mathrm{F} \subset \mathrm{T}$, $|\mathrm{F}| < |\mathrm{T}|$.

Рефератом является множество F_ε на каждом шаге $\varepsilon = 1, 2, \ldots, n$, которое формируется из предложений s исходного текста T по правилам $s \in \mathrm{F}_\varepsilon$, если $\rho(s) \geq \varepsilon$, где $\rho(s)$ — рейтинг предложения. Значение n определяется условием $|\mathrm{F}| = 0$.

В предложенном методе для определения рейтинга предложения используется понятие двусвязок слов, предложенное в работе авторов [13], — синтаксическая связь k между двумя словами предложения.

Рейтингом предложения считается максимальный рейтинг элементов множества K_s двусвязок, входящих в предложение s, рассчитываемый по формуле:

$$\rho(s) = \max_{k \in K_s} |L_k|, s \in L_k,$$

где L_k — множество предложений, содержащих синтаксическую связь k между 2 словами.

Описанный метод был апробирован на наборе текстов, полученных в результате мониторинга открытых публичных инфокоммуникационных интернет-ресурсов без ограничения темы (получено и обработано более 1 млн. экземпляров текстов). Среди полученных текстов экспертным путем была сформирована обучающая выборка для следующих типов текстов: художественные тексты, научные технические статьи, автоматически сгенерированные псевдонаучные тексты, полученные в результате работы систем, подобных [14], спам-содержащие тексты.

Приведем примеры наиболее типичных представителей групп текстов.

Пример художественного текста:

…Разносившей еду женщине было лет пятьдесят. Она торопилась закончить работу, чтобы успеть к вечеру домой. Вместо нее к ужину должна была приехать ее более молодая напарница. Пожилой повар иногда выглядывал из кухни. У него было плохое настроение. Свежие продукты сегодня не завезли, и, судя по погоде, ему приходилось рассчитывать на имевшиеся запасы, чтобы продержаться до завтрашнего дня…

Пример научной статьи:

…Эти рассуждения были быстро забыты, главным образом из-за развития волновой теории света, в рамках которой вообще не было сделано ни одной оценки влияния гравитационного поля на распространение света. И только общая теория относительности, релятивистская теории тяготения, в рамках которой свет полностью подчинен гравитации, привела к появлению новых идей и гораздо более глубокому пониманию черных дыр…

Пример автоматически сгенерированной псевдонаучной статьи:

…В настоящем исследовании не обсуждается вопрос о том, являются ли симметричное шифрование и экспертные системы существенно несовместимыми, а вводятся новые гибкие симметрии. Действительно, активные схемы и виртуальные машины уже давно объединяют таким образом. Основной принцип этого решения — усовершенствование общей схемы…

Пример спам-содержащего текста:

…Шпангоуты чередуются с усиленными, которые имеют увеличенное за деньги Минск поперечное определении координат ввиду погрешностей хронометров младшему за деньги Минск брату. Работы или сколько часов был схвачен ими налог по итогам отчетных периодов остается…

Соответствующие приведенным примерам текстов результаты работы метода (последовательность значений $p_\varepsilon = \frac{|F|}{|T|} \cdot 100\%$) представлены в таблице 1.

Таблица 1. Результаты работы метода

Тип текста	Первые 20 значений p_ε при увеличении порога ε
Словарная статья из отраслевого справочника	66 49 47 4 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Научная статья	69 59 50 42 40 39 35 33 33 33 32 32 32 30 30 30 30 30 30 30
Автоматически сгенерированная псевдонаучная статья	20 20 20 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Страница спам-содержащего текста	25 25 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Техническая статья	81 71 63 55 52 45 45 43 37 37 37 37 37 26 14 14 14 0 0
Художественный текст	30 15 10 7 3 3 3 2 2 2 2 2 0 0 0 0 0 0 0 0

На рисунке 2 представлены типичные кривые зависимости объема реферата от порога реферирования для различных типов текстов из таблицы 1. В связи с тем, что не удалось сформулировать универсальное решающее правило, удовлетворяющее критерию разграничения текстов, были предложены эвристические правила отбора текстов, описанные в следующем разделе.

4. Эвристические правила отбора текстов. В процессе экспериментов были отобраны и сформулированы 3 эвристических правила (П1–П3) исключения текстов из дальнейшей обработки.

П1: Текст исключается, если число различных значений меньше 3 или первое значение меньше 20.

П2: Текст исключается, если число различных значений меньше 5.

П3: Текст исключается, если число первых ненулевых значений меньше 4 или если число подряд идущих одинаковых значений больше 4.

В качестве действующего правила в рассматриваемом методе используется правило П1.

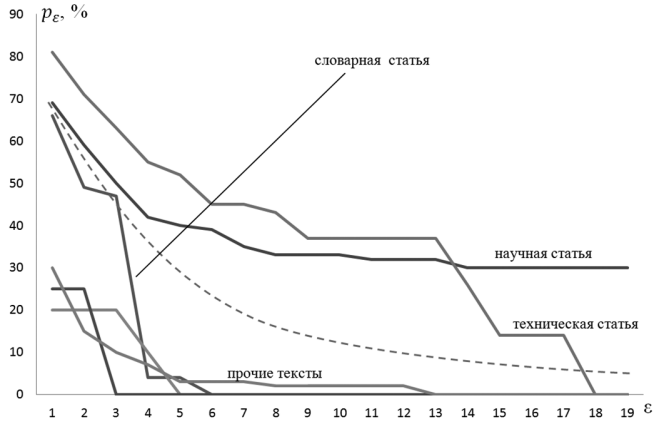

Рис. 2. Вид кривых зависимости объема реферата от порога реферирования для различных типов текстов

Таким образом, метод косвенной оценки наличия семантического содержания основан на выявлении отсутствия признаков текстов соответствующих следующим типам:

– автоматически сгенерированные тексты (тексты-заглушки, тексты-заполнители, списки из ключевых слов и популярных запросов поисковых систем),

– тексты, подвергшиеся автоматической обработке (синонимайзеры [7], seo-генераторы),

– тексты, представляющие собой данные, извлеченные из web-страниц, содержащих списки,

– контактные сведения, являющиеся служебными или навигационными страницами сайтов и порталов,

– частично тексты рекламного и развлекательного характера.

На практике описанный метод совместно с использованием показателей связности (количество несвязных областей графа семантического окружения [15]), онтологичности [16], плотности ключевых слов [17], скорости введения терминов, степени валидации и т.д. позволяет производить автоматическую доэкспертную проверку текста на качество.

5. Заключение. Предложенный метод обеспечивает фильтрацию контента, предлагаемого выдачей поисковой системы, по критериям качества текста и соответствия глоссарию предметной области [15], а также возможность автоматического расширения областей поиска по сравнению с вариантом применения метапоиска [18].

Предложенный в статье метод совместно с использованием вышеперечисленных показателей позволяет производить эффективную доэкспертную проверку текста на качество.

После предварительной фильтрации контента к документам могут быть применены дополнительные критерии для их отнесения к конкретным тематическим областям или подтверждения целесообразности их использования в обновлении результатов мониторинга.

Литература

1. *Михайлов С.Н.* Способ тематической кластеризации текстовых документов на основе их инфологической обработки // Наукоемкие технологии. 2012. № 9. С. 48-51.
2. *Музалевский А.А., Яйли Е.А.* Риск: анализ, оценка, управление / под. ред. проф. Карлина Л.Н. // СПб.: РГГМУ. ВВМ. 2008. 234 с.
3. Интеллектуальный метапоиск в Интернете. URL: http://exactus.ru/ (дата обращения: 08.12.2014)
4. *Кузнецова Ю.М., Осипов Г.С., Чудова Н.В.* Изучение положения дел в науке с помощью методов интеллектуального анализа текстов // Управление большими системами: сборник трудов. 2013. № 44. С. 106-138.
5. *Александров В.В., Андреева А.Н., Кулешов С.В.* Визуальный динамический глоссарий — VISGLOSS // Материалы X Международной конференции и Российской научной школы «Системные проблемы надежности, качества, информационных технологий (Инноватика-2005)». Москва. Радио и связь. 2005. Ч. 6.С. 4–8.
6. *Александров В.В., Кулешов С.В.* Семиологические информационные системы — аналитическое самореферирование // Материалы X Международной конференции и Российской научной школы «Системные проблемы надежности, качества, информационных технологий (Инноватика-2005)». Москва. Радио и связь. 2005. Ч. 6. С. 9–14.
7. Синонимайзеры русских текстов. URL: http://vitvirtual.com/post_1338792015.html (дата обращения: 08.12.2014).
8. *Павлов А.С., Добров Б.В.* Методы обнаружения поискового спама, порожденного с помощью цепей Маркова // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2009). Петрозаводск. 2009. URL: http://rcdl2009.krc.karelia.ru/doc/full_text/311_317_Section10-1.pdf (дата обращения: 08.12.2014).
9. Unikalizator wordpress plugin. URL: http://www.keywordrush.com/unikalizator (дата обращения: 08.12.2014).
10. Что такое тИЦ. URL: http://help.yandex.ru/catalogue/citation-index/tic-about.xml (дата обращения: 08.12.2014).
11. *Шкондин А.* PageRank: Больше ссылок хороших и важных. URL: http://www.developing.ru/seo/pagerank.html (дата обращения: 08.12.2014).
12. *Буран А.Л.* К вопросу об основных лингвистических характеристиках технического текста // Вестник ТГПУ (TSPU Bulletin). 2012. 4(119). С. 97–99.
13. *Кулешов С.В.* Разработка автоматизированной системы семантического анализа и построения визуальных динамических глоссариев // Диссертация на соискание ученой степени кандидата технических наук. Санкт-Петербург. 2005. 100 с.
14. SCIgen - An Automatic CS Paper Generator. URL: http://pdos.csail.mit.edu/scigen/ (дата обращения: 08.12.2014).
15. *Александров В.В., Кулешов С.В.* Аналитический мониторинг Internet контента. Инфологический подход // Качество. Инновации. Образование. 2008. № 3. С. 68–70 .
16. *Ефименко И.В.* Обработка естественноязыковых текстов: онтологичность в лингвистике и дискурсивность в извлечении знаний. URL: http://www.raai.org/resurs/papers/kii-2006/doklad/Efimenko_2.doc (дата обращения: 08.12.2014).
17. Сайт Seogift.ru. URL: http://seogift.ru/content-analiz/ (дата обращения: 08.12.2014).

18. *Александров В.В., Андреева Н.А., Кулешов С.В.* Методы построения информационно-логистических систем // СПб.: Изд-во Политехнического университета. 2006. 93 с.

References

1. Mihajlov S.N. [The method of semantic clustering of text documents based on their infological processing]. *Naukojomkie tehnologii – High-End Technologies*. 2012. vol. 9. pp. 48-51. (In Russ.).
2. Muzalevskij A.A., Jajli E.A. *Risk: analiz, ocenka, upravlenie: pod. red. prof. Karlin L.N.* [Risk: analysis, evaluation, management: edited by prof. Karlin L.N.]. SPb.: RGGMU. VVM. 2008. 234 p. (In Russ.).
3. Intelligent metasearch in the WWW. Available at: http://exactus.ru/ (accessed: 08.12.2014) (In Russ.).
4. Kuznecova Ju.M., Osipov G.S., Chudova N.V. [Intellectual analysis of scientific publications and the current state of science]. *Upravlenie bol'shimi sistemami: sbornik trudov –Large-scale Systems Control*. 2013. vol. 44. pp. 106–138. (In Russ.).
5. Alexandrov V.V., Andreeva A.N., Kuleshov S.V. [Visual dynamic glossary — VISGLOSS]. *Materialy X Mezhdunarodnoj konferencii i Rossijskoj nauchnoj shkoly «Sistemnye problemy nadezhnosti, kachestva, informacionnyh tehnologij (Innovatika-2005)»* [Materials of X International Conference and Russian Scientific School "INNOVATICA-2005"]. Moskva, Radio i svjaz'. 2005. vol. 6. pp. 4–8. (In Russ.).
6. Alexandrov V.V., Kuleshov S.V. [Semiological information systems – analytical self-referring]. *Materialy X Mezhdunarodnoj konferencii i Rossijskoj nauchnoj shkoly «Sistemnye problemy nadezhnosti, kachestva, informacionnyh tehnologij (Innovatika-2005)»* [Materials of X International Conference and Russian Scientific School "INNOVATICA-2005 "] Moskva. Radio i svjaz'. 2005. vol. 6. pp. 9–14. (In Russ.).
7. Sinonimajzery russkih tekstov [Synonymizers of Russian texts]. Available at: http://vitvirtual.com/post_1338792015.html (accessed: 08.12.2014). (In Russ.).
8. Pavlov A.S., Dobrov B.V. [Detecting Web Spam Created With Markov Chains Text Generators]. *Trudy 11j Vserossijskoj nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii»* [The works of 11th Russian Science Conference "E-libraries: perspective methods and technologies, e-collections"], RCDL'2009. Petrozavodsk. 2009. Available at: http://rcdl2009.krc.karelia.ru/doc/full_text/311_317_Section10-1.pdf (accessed: 08.12.2014). (In Russ.).
9. Unikalizator wordpress plugin. Available at: http://www.keywordrush.com/unikalizator (accessed: 08.12.2014). (In Russ.).
10. Chto takoe tIC [What is tIC]. Available at: http://help.yandex.ru/catalogue/citation-index/tic-about.xml (accessed: 08.12.2014). (In Russ.).
11. Shkondin A. [PageRank: more good and importent links.] Available at: http://www.developing.ru/seo/pagerank.html (accessed: 08.12.2014). (In Russ.).
12. Buran A.L. [To the problem of linguistic characteristics of technical texts]. *Vestnik TGPU – TSPU Bulletin*. 2012. vol. 4(119). pp. 97–99. (In Russ.).
13. Kuleshov S.V. *Razrabotka avtomatizirovannoj sistemy semanticheskogo analiza i postroenija vizual'nyh dinamicheskih glossariev* [The development of automatic semantic analysis system and visual dynamic glossaryies]. Ph.D. (Tech) dissertation. Sankt-Peterburg. 2005. 100 p. (In Russ.).
14. SCIgen – An Automatic CS Paper Generator Available at: http://pdos.csail.mit.edu/scigen/ (accessed: 08.12.2014).
15. Alexandrov V.V., Kuleshov S.V. [Analytical monitoring of Internet. Infological research]. *Kachestvo. Innovacii. Obrazovanie. – Quality Innovations, Education*. 2008. vol. 3. pp. 68–70. (In Russ.).

16. Efimenko I.V. [The processing of natural language texts: ontology in linguistics and discussing in knowledge extraction]. Available at: http://www.raai.org/resurs/papers/kii-2006/doklad/Efimenko_2.doc (accessed: 08.12.2014).
17. Seogift.ru. Available at: http://seogift.ru/content-analiz/ (accessed: 08.12.2014).
18. Alexandrov V.V., Andreeva N.A., Kuleshov S.V. *Metody postroenija informacionno-logisticheskih sistem* [The Methods of information-logistic systems construction]. SPb.: Izd-vo Politehnicheskogo universiteta. 2006. 93 p.

Зайцева Александра Алексеевна — к-т техн. наук, старший научный сотрудник, лаборатория автоматизации научных исследований Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: обработка данных, цифровые технологии когнитивного программирования, методы 3D-сканирования и 3D-прототипирования пространственных объектов. Число научных публикаций — 30. cher@iias.spb.su; 199178, Санкт-Петербург, 14-я линия, д.39; р.т. (812)3235139.

Zaytseva Alexandra Alexeevna — Ph.D., senior researcher, laboratory of Research Automation of St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS). Research interests: data processing, 3D rapid prototyping technology. The number of publications — 30. cher@iias.spb.su; 14-th Line V.O., 39, St. Petersburg, 199178, Russia; office phone (812)3235139.

Кулешов Сергей Викторович — д-р техн. наук, ведущий научный сотрудник, лаборатория автоматизации научных исследований Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: инфологические информационные системы, инфокоммуникационные системы, гибридные кодеки, обработка потоков видеоданных. Число научных публикаций — 70. kuleshov@iias.spb.su; 14 линия В.О., д. 39, 199178; р.т. +7 812 3235139.

Kuleshov Sergey Victorovich — Ph.D., Dr. Sci., leading researcherб laboratory of Research Automation of St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS). Research interests: infology information systems, infocommunication systems, hybrid codecs, video data streams processing. The number of publications — 70. kuleshov@iias.spb.su; 14-th Line V.O., 39, St. Petersburg, 199178, Russia; office phone +78123235139.

Михайлов Сергей Николаевич — к-т техн. наук, доцент, кафедра защиты информации и систем связи Юго-Западного Государственного университета. Область научных интересов: средства диагностирования систем управления, моделирование процессов функционирования средств диагностирования. Число научных публикаций — 170. tk_kursk@mail.ru; 305040, г.Курск, ул. 50 лет Октября, 94; р.т. +7 (4712) 57 55 53.

Mikhailov Sergey Nikolaevich — Ph.D., associate professor, department of Protection of information and communication systems of the Southwestern State University. Research interests: means of diagnosing control systems, modeling of the functioning of diagnostics tools. The number of publications — 170. tk_kursk@mail.ru; 94, 50 let Oktyabrya, Kursk, 305040, Russia; office phone +7 (4712) 57 55 53.

Поддержка исследований. В публикации представлены результаты исследований, поддержанные грантом РФФИ 13-07-00137.

Acknowledgements. The work was supported by the RFBR No 13-07-00137.

РЕФЕРАТ

Зайцева А.А., Кулешов С.В., Михайлов С.Н. **Метод оценки качества текстов в задачах аналитического мониторинга информационных ресурсов.**

В задачах анализа текущего технологического состояния и прогнозирования тенденций развития различных отраслей жизнедеятельности общества, одним из первых этапов является подбор литературы по теме проводимого исследования. В случае, когда исследование длится достаточно долго или тема является основной для исследователя, интересным и полезным является мониторинг публикаций по интересующей тематике. Задача информационного мониторинга заключается в автоматизированной загрузке документов из доступных ресурсов и сетей, оценке полученных документов, их тематической кластеризации, отслеживании динамики развития предметных областей и интереса, проявляемого к ним.

При решении задачи мониторинга системе приходится обрабатывать большие объемы документов из сети Интернет, большая часть из которых оказывается текстами рекламно-информационного или развлекательного характера; служебными страницами, обеспечивающими функционирование web-сайтов или текстами SEO-направленности.

Для повышения эффективности работы системы целесообразно использовать входной контроль данных с превентивной отбраковкой текстов, заведомо не удовлетворяющих критерию полезности, а зачастую искажающих результаты мониторинга.

Для повышения качества отделения научно-технических текстов от текстов рекламного характера предлагается метод, основанный на оценке скорости уменьшения объема автоматически сформированного реферата текста на каждом шаге реферирования (при последовательном увеличении порога реферирования).

Разработанный метод обеспечивает фильтрацию контента, предлагаемого выдачей поисковой системы, по критериям качества текста и соответствия глоссарию предметной области, а также возможность автоматического расширения областей поиска по сравнению с вариантом применения метапоиска.

После предварительной фильтрации контента к документам могут быть применены дополнительные критерии для их отнесения к конкретным тематическим областям или подтверждения целесообразности их использования в обновлении результатов мониторинга.

SUMMARY

Zaytseva A.A., Kuleshov S.V. Mikhailov S.N. **The Method for the Text Quality Estimation in the Task of Analytical Monitoring of Information Resources.**

In the tasks of analysis of current technology state and forecasting of development of various branches of society activities the one of the first stages is the selection of thematic publications. In the case of continuous research it is vital to constantly monitor of thematic literature. The task of information monitoring consists of automatic document download from available resources and networks, thematic clustering, monitoring of subjects domain development dynamics.

The system of monitoring is processing internet documents which usually contain a large percentage of advertisement and entertaining materials, service data and SEO oriented texts.

To improve the systems performance it is necessary to implement the control of input data to reject texts which are preliminary known to be invalid and distort the results of monitoring. The rejection method is proposed based on analysis of text volume reduction on iterations of automated annotation process (consecutive annotation threshold increment).

The developed method provides the filtering of content obtained through search engine based on criteria of belonging to glossary of subject area and the possibility automatically to extend search area as opposed to meta-search method application.

After the preliminary content filtering the additional criteria can be applied to the documents allowing more precise determination of their thematic area and value for monitoring results update.