3. *Klimenko I.S., Korovko P.G., Sharapova L.V.* K probleme ocenivaniya ehffektivnosti upravleniya i kachestva upravlencheskikh reshenij // Vestnik Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2017. № 1. S. 53–57.
4. *Klimenko I.S., Sharapova L.V.* K issledovaniyu fenomena informacii // Vestnik Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2014. № 4. S. 141–149.
5. *Klimenko I.S., Sharapova L.V.* K probleme sistemnogo analiza telekommunikacionnykh processov // Rossijskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2016. № 1–2. S. 82–86.
6. *Kharkevich A.A.* O cennosti informacii // Problemy kibernetiki. M.: Fizmatgiz, 1960. Vyp. 4.
7. *Shennon K.* Raboty po teorii informacii i kibernetike. M.: Fizmatgiz, 1959.

DOI: 10.25586/RNU.V9187.19.02.P.157 УДК 004.912+002.513.5

С.А. Краснов

О ВОЗМОЖНОСТИ СМЫСЛОВОГО АНАЛИЗА ИНФОРМАЦИИ ДЛЯ ВЫЯВЛЕНИЯ ИНФОРМАЦИОННЫХ ИНТЕРЕСОВ ПОЛЬЗОВАТЕЛЕЙ

Рассматривается возможность применения метода латентно-семантического анализа (ЛСА) для выявления информационных интересов пользователей в web-пространстве. Показана положительная динамика применения метода ЛСА в различных направлениях, где требуется смысловой анализ информации. Предложены основные модули и схема реализации программного комплекса, позволяющего осуществлять сбор, обработку и представление информационных интересов пользователей.
Ключевые слова: латентно-семантический анализ, смысловой анализ информации, рубрикация, информационные интересы пользователей.

S.A. Krasnov

ABOUT THE POSSIBILITY OF SEMANTIC ANALYSIS OF INFORMATION TO IDENTIFY INFORMATIONAL INTERESTS OF USERS

The possibility of using the method of latent-semantic analysis (LSA) to identify informational interests of users in the web space is considered. The positive dynamics of applying the LSA method in different directions, where a semantic analysis of information is required, is shown. The main modules and scheme for the implementation of the software package that allows for the collection, processing and presentation of information interests of users are proposed.
Keywords: latent-semantic analysis, semantic analysis of information, rubrication, informational interests of users.

Одной из важных и неотъемлемых частей информационного обеспечения различных организаций являются процессы поиска, отбора и рубрикации информации, полученной из пространства глобальной сети Интернет [4]. В настоящее время особенно остро сто-

ят вопросы поиска, отбора и автоматического смыслового рубрицирования информации в условиях динамичного изменения экономико-политической обстановки и ее факторов, когда многократно вырос перечень информационных агентств и порталов, объемы передаваемого ими контента. При этом перечень задач по информационному обеспечению руководства организаций и других заинтересованных сотрудников постоянно растет.

В настоящее время, несмотря на проведение работ по исследованию и разработке методов и программного обеспечения поиска, анализа и структурирования информации, их применение носит инициативный и разрозненный характер с задействованием прежде всего доступных коммерческих систем, которые не настроены на решение специфических задач поиска, отбора и рубрикации информации, востребованной пользователями. Кроме того, они реализованы на устаревших статистических методах и алгоритмах.

Сложившаяся ситуация затрудняет процессы мониторинга и анализа информационных интересов пользователей глобальной сети Интернет, ее автоматическое смысловое рубрицирование, необходимое для удобства анализа и включения полученной информации в процессы, направленные на выработку решений для адаптации организации к современным востребованным условиям.

Исходя из вышесказанного необходимо разработать программный комплекс, основанный на современных методах интеллектуального смыслового анализа информации, позволяющий структурировать получаемую информацию из глобальной сети Интернет в требуемом представлении. Ее предназначение заключается в непрерывном мониторинге и обобщении больших объемов текстовой информации, полученной из глобальной сети Интернет, в процессе наблюдения за одиночными информационными потребностями конкретного пользователя или группы пользователей.

Смысловое структурирование предлагается осуществлять на базе методов и методик смысловой обработки информации, принципиально отличающихся от доминирующих в настоящее время статистических методов, выявлением скрытых семантических взаимосвязей между информационными признаками всего множества неструктурированных данных с помощью метода ЛСА [3; 10].

Основная идея латентно-семантического анализа (ЛСА) заключается в том, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые позволяют определить похожесть смысловых значений слов и множеств слов между собой. Кроме того, ЛСА измеряет корреляционные зависимости типа «терм – терм», «терм – вектор» и «вектор – вектор». Результативность данного метода зависит не только от частот использования слов (термов) в документах, но и от выявления более глубоких (скрытых) связей.

Исходной информацией для ЛСА является матрица термов на документы, которая описывает используемый для обучения системы набор данных. Элементы этой матрицы содержат частоты использования каждого терма в каждом документе.

Один из самых распространенных вариантов ЛСА основан на использовании разложения исходной матрицы по сингулярным значениям (SVD – Singular-Value Decomposition). Используя SVD, большая исходная матрица разлагается во множество из k, обычно от 70 до 200 ортогональных матриц, линейная комбинация которых является хорошим приближением исходной матрицы.

Согласно теореме о сингулярном разложении, любая вещественная прямоугольная матрица X может быть разложена в произведение трех матриц:

$$X = U\Sigma V^T,$$

где матрицы U и V – ортогональные, а Σ – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы X.

Особенность такого разложения в том, что если в Σ оставить только k наибольших сингулярных значений, а в матрицах U и V – только соответствующие этим значениям столбцы, то произведение получившихся матриц Σ_{lsa}, U_{lsa} и V_{lsa} будет наилучшим приближением исходной матрицы X матрицей ранга k:

$$X \cong \hat{X} = U_{lsa}\Sigma_{lsa}V_{lsa}.$$

Идея такого разложения и суть латентно-семантического анализа заключается в том, что если в качестве X использовалась матрица термов на документ, то матрица $\hat{X}$, содержащая только k первых линейно независимых компонент X, отражает основную структуру ассоциативных зависимостей, присутствующих в исходной матрице, и в то же время не содержит шума.

Таким образом, каждый терм и документ представляются при помощи векторов в общем пространстве размерности k (так называемом *пространстве гипотез*). Близость между любой комбинацией термов или документов может быть легко вычислена при помощи скалярного произведения векторов.

Метод анализа динамики изменения сингулярных чисел матрицы «терм – документ» с автоматическим выбором диапазона используемых ранговых значений.

На первом шаге необходимо сформировать матрицу А «терм – документ», которая опишет анализируемые документы и будет содержать исходные данные для метода ЛСА. Ее элементы будут содержать веса термов, полученные после применения статистической меры *tf – idf* (отношение частоты слов к инверсной частоте документа) $0 \le d_{ij} \le 1$ и НВ матрицы:

$$d_{ij} = \frac{w_{ij}}{\left|\overrightarrow{w_{ij}}\right|}, \quad w_{ij} = tf_{ij}\log\frac{|D|}{df_j}, \quad tf_{ij} = \frac{n_i}{\sum_k n_k}, \tag{1}$$

где d_{ij} – нормированные веса по *tf – idf*, (частота встречаемости термина – обратная документная частота), $0 \le d_{ij} \le 1$;

$\left|\overrightarrow{w_{ij}}\right|$ – нормированный вектор w_{ij} в евклидовом пространстве;

df_j – документная частота (число документов, в которых встретилось j-е слово);

$|D|$ – число анализируемых документов;

n_i – количество употреблений слова в документе;

$\Sigma_k n_k$ – общее количество слов содержащихся в документе;

tf_{ij} – частота встречаемости слова в документе (число раз, которое j-е слово встретилось в i-м документе).

В *tf – idf* наибольший вес получают слова с высокой частотой в пределах документа и с небольшой частотой встречаемости в других документах. При НВ матрицы «терм – документ» скалярное произведение не зависит от нормы векторов. Это позволяет упростить сравнение результатов скалярных произведений. Операция нормирования производится перед расчетом l – степени соответствия документов.

Далее необходимо выполнить сингулярное разложение матрицы A в произведение трех матриц:

$$A = UWV^T, \quad (2)$$

где U и V – унитарные матрицы, которые состоят из левых и правых сингулярных векторов, а W – матрица с неотрицательными элементами на диагонали, которые называются сингулярными значениями матрицы A.

Согласно теореме Эккарта – Янга, если в матрице W оставить только наибольшие сингулярные значения σ, а в матрицах U и V – соответствующие этим значениям столбцы, то матрицы U_σ и V_σ будут лучшими их приближениями, отражающими ассоциативные зависимости представления термов и документов в пространстве размерности σ [17; 20].

Следующим шагом является оптимальный выбор ненулевых сингулярных значений $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n > 0$ матрицы A, которые влияют на результат выявления противоречивой и дублирующей информации. Результатом приведения матрицы A к рангам, имеющим близкие к нулю сингулярные значения, являются равные матрицы, учет которых ведет к увеличению вычислительной сложности метода ЛСА и тем самым снижает оперативность выявления противоречивой и дублирующей информации. Чтобы решить задачу оптимального выбора сингулярных значений предлагается описанный ниже эвристический метод выбора значимых рангов, который является сущностью метода анализа динамики изменения сингулярных чисел матрицы «терм –документ» с автоматическим выбором диапазона используемых ранговых значений.

Определим функцию $f(i) = \sigma_i$, $i \in N$, $i < P$, где P – количество документов. Значимыми рангами являются только ранги r_p, r_{p+1}, ..., r_{m-1}, r_m, $p \leq m$, заключенные между соответствующими сингулярными значениями $\sigma_p \geq \sigma_m \geq 0$, претерпевающими резкое изменение $\Delta\sigma_i = \sigma_i - \sigma_{i-1}$, $i \in \{p; m\}$ относительно предыдущих сингулярных значений $\sigma_i \geq \sigma_p$, $i \leq p$; $\sigma_i \geq \sigma_m$, $i \leq m$.

Определение границ значимых рангов осуществляется с помощью понятия производной функции сингулярных значений $f'(i) = \sigma_i - \sigma_{i-1}$, $i \in N$, $i < n$.

Далее осуществляется поиск максимального значения производной функции $f'_{\max}$, достижимого при $\sigma_{\max}$, $1 < \max \leq \frac{n}{2}$.

Затем определяется первый локальный минимум σ_p, следующий за $\sigma_{\max}$. Ранги r_1, r_2, ..., r_i, ..., r_{p-1}, r_p, $i \leq p$, соответствующие сингулярным значениям, большим σ_p признаются незначимыми.

В качестве правой границы значимых рангов выбирается ранг r_n, соответствующий последнему ненулевому сингулярному значению σ_n.

На заключительном этапе необходимо рассчитать λ документов, используя КМБ:

$$\cos(\vec{X}_j, \vec{X}_i) = \sum\nolimits_{i=1}^{M} x_j^{(i)} x_i^{(i)}, \quad (3)$$

где $x_j^{(i)} x_i^{(i)}$ – элементы разных векторов, между которыми вычисляется мера близости; M – размерность пространства векторов.

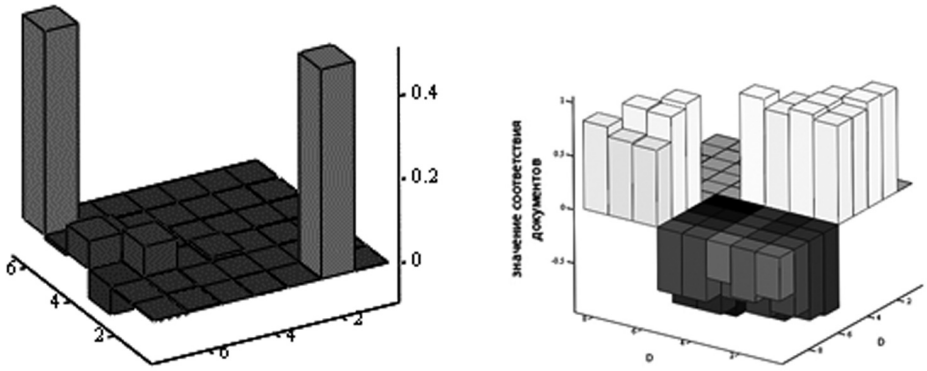
Значения КМБ ограничены промежутком [–1; 1] при использовании операции НВ. Степень соответствия $\lambda_{j,i}^{l}$ векторов $\vec{X}_j, \vec{X}_i$ $(i < j \leq P)$ вычисляется для каждого значимого ранга r_l, $p \leq l \leq m$. Далее необходимо вычислить результирующую $\overline{\lambda}_{i,j}$ векторов $\vec{X}_j, \vec{X}_i$:

$$\overline{\lambda}_{j,l} = \frac{\sum\nolimits_{l=p}^{m} \lambda_{j,i}^{l}}{m - p + 1}. \quad (4)$$

Таким образом, получаем результирующую степень соответствия для конкретной пары документов по всему значимому диапазону ранговых значений. Полученное значение необходимо сравнить, например, с пороговым значением для автоматического или автоматизированного принятия решения по устранению дублирующих и (или) противоречивых данных, классификации информации и (или) фильтрации информации.

Отметим, что метод автоматического определения диапазона используемых ранговых значений позволяет с большей точностью гарантировать, что данные действительно противоречивые или дублирующие, потому что значения l всех пар векторов оцениваются на каждом значимом ранге. При этом случайные всплески полученных значений l при неправильных ранговых значениях сглаживаются, а значения l явно дублирующих (противоречивых) данных стремятся к единице, так как находится среднее арифметическое по всем значениям l одной пары векторов матриц, полученных из диапазона используемых ранговых значений.

Результаты исследований показали, что применение метода ЛСА в задачах поиска, рубрикации, выявления дублирующей информации позволяет эффективно выявлять смысловые взаимосвязи между термами. Это позволило повысить автоматизацию решения задачи устранения конфликтов и избыточности информации, повысить точность при поиске рубрикации информации. Гистограммы, представленные на рисунке, это подтверждают. Гистограмма на рисунке *а* указывает на меры близости различных групп документов, а на рисунке *б* – на дублирование заголовков [2; 6; 7; 11].



а *б*

Гистограммы, показывающие эффективность применения метода ЛСА

Исходя из вышесказанного можно предположить, что для решения задачи выявления информационных интересов пользователей в глобальной сети Интернет будет эффективен метод ЛСА. Поэтому разрабатываемый ПК должен включать в себя модули:

- слежения за определенным диапазоном в истории посещения информационных ресурсов web-пространства;
- автоматического сбора информации;
- анализа, кластеризации и рубрикации информации на основе метода ЛСА;
- настраиваемых отчетных документов и форм.

Автоматизация и непрерывность процессов сбора и рубрикации информации на базе интеллектуальных методов и методик ее обработки (лексического, синтаксического и се-

мантического анализа) позволят повысить оперативность и точность рубрикации информации [1; 5; 7; 8; 9].

Для обеспечения процесса выявления информационных интересов пользователей необходимо:

1) выбрать определенный диапазон в истории посещения информационных ресурсов web-пространства отобранным пользователем или группой пользователей, т.е. отследить текущую или несколько сессий его или их работы;

2) последовательно пройтись по истории сессий, осуществить автоматический сбор текстовой информации с каждого информационного ресурса;

3) произвести кластеризацию с последующей рубрикацией полученной информации посещенных им или ими web-страниц при помощи методов смыслового анализа информации;

4) проанализировать результаты рубрикации;

5) сделать вывод об информационных интересах контролируемого пользователя или группы пользователей;

6) Выдать рекомендаций сотрудникам организации, корректирующие функционирование их деятельности с учетом изменения потребностей пользователей.

Применение предложенной схемы выявления потребностей пользователя позволит организации подстраиваться под их интересы своевременно. Это позволит организации находиться на ведущих позициях и своевременно подстраиваться под динамически изменяемую экономическую и политическую ситуацию как в стране, так и в мире.

Литература

1. *Войцеховский С.В., Калиниченко С.В., Краснов С.А., Уланов А.В.* Модель оценивания оперативности обработки устаревающей информации // Научное обозрение. 2014. № 3. С. 155–157.

2. *Краснов С.А., Илатовский А.С., Хомоненко А.Д., Арсеньев В.Н.* Оценка семантической близости документов на основе латентно-семантического анализа с автоматическим выбором ранговых значений // Труды СПИИРАН. 2017. № 5 (54). С. 185–204.

3. *Краснов С.А.* Математическая модель метода латентно-семантического анализа в системе семантической рубрикации документов // Компьютерные технологии и информационные системы: сб. науч. тр. ВА ВПВО ВС РФ. Смоленск, 2011. Вып. 18. С. 33–43.

4. *Краснов С.А.* Обзор моделей поиска и методов тематического анализа текстовой информации // Компьютерные технологии и информационные системы: сб. науч. тр. ВА ВПВО ВС РФ. 2011. Вып. 20. С. 35–42.

5. *Краснов С.А. Уланов А.В, Матвеев С.В.* Анализ оперативности обработки информации с ограниченным временем актуальности // Бюллетень результатов научных исследований: электрон. науч. журн. ПГУПС. 2013. Вып. 9 (4). С. 39–47.

6. *Краснов С.А., Хомоненко А.Д., Дашонок В.Л.* Выявление противоречий в семантически близкой информации на основе латентно-семантического анализа // Проблемы информационной безопасности. Компьютерные системы: сб. науч. тр. СПбГПУ. 2014. № 2. С. 73–84.

7. *Краснов С.А., Хомоненко А.Д., Яковлев Я.В.* Оценка эффективности применения алгоритма вычисления коэффициента ранговой корреляции Спирмена в методе латентно-семантического анализа при семантической рубрикации документов // Электронный научный журнал «Бюллетень результатов научных исследований». 2012. Вып. 3(2). С. 153–162.

8. *Хомоненко А.Д. Бубнов В.П., Краснов С.А., Еремин А.С.* Модель функционирования системы автоматической рубрикации документов в нестационарном режиме // Проблемы информационной безопасности. Компьютерные системы. 2011. № 4. С. 16–23.

9. *Хомоненко А.Д., Краснов С.А., Еремин А.С.* Оценка оперативности автоматической рубрикации документов с помощью модели нестационарной системы обслуживания с эрланговским распределением длительности интервалов между запросами // Проблемы информационной безопасности. Компьютерные системы. 2012. № 3. С. 14–21.
10. *Хомоненко А.Д., Краснов С.А.* Применение метода латентно-семантического анализа для автоматической рубрикации текстов в системах электронного документооборота // Сборник материалов первой международной научно-практической конференции. СПб.: ПГУПС, 2011. С. 291–294.
11. *Хомоненко А.Д. Краснов С.А.* Применение методов латентно-семантического анализа для автоматической рубрикации документов // Известия Петербургского университета путей сообщения. 2012. Вып. 2 (31). С. 124–132.

Literatura

1. *Vojcekhovskij S.V., Kalinichenko S.V., Krasnov S.A., Ulanov A.V.* Model' ocenivaniya operativnosti obrabotki ustarevayushchej informacii // Nauchnoe obozrenie. 2014. № 3. S. 155–157.
2. *Krasnov S.A., Ilatovskij A.S., Khomonenko A.D., Arsen'ev V.N.* Ocenka semanticheskoj blizosti dokumentov na osnove latentno-semanticheskogo analiza s avtomaticheskim vyborom rangovykh znachenij // Trudy SPIIRAN. 2017. № 5 (54). S. 185–204.
3. *Krasnov S.A.* Matematicheskaya model' metoda latentno-semanticheskogo analiza v sisteme semanticheskoj rubrikacii dokumentov // Komp'yuternye tekhnologii i informacionnye sistemy: sb. nauch. tr. VA VPVO VS RF. Smolensk, 2011. Vyp. 18. S. 33–43.
4. *Krasnov S.A.* Obzor modelej poiska i metodov tematicheskogo analiza tekstovoj informacii // Komp'yuternye tekhnologii i informacionnye sistemy: sb. nauch. tr. VA VPVO VS RF. 2011. Vyp. 20. C. 35–42.
5. *Krasnov S.A. Ulanov A.V, Matveev S.V.* Analiz operativnosti obrabotki informacii s ogranichennym vremenem aktual'nosti // Byulleten' rezul'tatov nauchnykh issledovanij: ehlektron. nauch. zhurn. PGUPS. 2013. Vyp. 9 (4). S. 39–47.
6. *Krasnov S.A., Khomonenko A.D., Dashonok V.L.* Vyyavlenie protivorechij v semanticheski blizkoj informacii na osnove latentno-semanticheskogo analiza // Problemy informacionnoj bezopasnosti. Komp'yuternye sistemy: sb. nauch. tr. SPbGPU. 2014. № 2. S. 73–84.
7. *Krasnov S.A., Khomonenko A.D., Yakovlev Ya.V.* Ocenka ehffektivnosti primeneniya algoritma vychisleniya koehfficienta rangovoj korrelyacii Spirmena v metode latentno-semanticheskogo analiza pri semanticheskoj rubrikacii dokumentov // Ehlektronnyj nauchnyj zhurnal "Byulleten' rezul'tatov nauchnykh issledovanij". 2012. Vyp. 3 (2). S. 153–162.
8. *Khomonenko A.D. Bubnov V.P., Krasnov S.A., Eremin A.S.* Model' funkcionirovaniya sistemy avtomaticheskoj rubrikacii dokumentov v nestacionarnom rezhime // Problemy informacionnoj bezopasnosti. Komp'yuternye sistemy. 2011. № 4. S. 16–23.
9. *Khomonenko A.D., Krasnov S.A., Eremin A.S.* Ocenka operativnosti avtomaticheskoj rubrikacii dokumentov s pomoshch'yu modeli nestacionarnoj sistemy obsluzhivaniya s ehrlangovskim raspredeleniem dlitel'nosti intervalov mezhdu zaprosami // Problemy informacionnoj bezopasnosti. Komp'yuternye sistemy. 2012. № 3. S. 14–21.
10. *Khomonenko A.D., Krasnov S.A.* Primenenie metoda latentno-semanticheskogo analiza dlya avtomaticheskoj rubrikacii tekstov v sistemakh ehlektronnogo dokumentooborota // Sbornik materialov pervoj mezhdunarodnoj nauchno-prakticheskoj konferencii. SPb.: PGUPS, 2011. S. 291–294.
11. *Khomonenko A.D. Krasnov S.A.* Primenenie metodov latentno-semanticheskogo analiza dlya avtomaticheskoj rubrikacii dokumentov // Izvestiya Peterburgskogo universiteta putej soobshcheniya. 2012. Vyp. 2 (31). S. 124–132.

РЕКОМЕНДАЦИИ АВТОРАМ

1. Статьи, направляемые для публикации в журнал, должны освещать результаты исследований и/или практический опыт и содержать информацию, открытую для печати и представляющую научный и практический интерес. Статьи аспирантов, докторантов, соискателей ученой степени, указываемые в списках научных трудов, как правило, должны отражать основные результаты их диссертационных исследований.

Статьи представляются на русском или английском языках.

Объем статьи должен составлять 12 000–18 000 знаков с пробелами (включая аннотацию и список литературы).

2. В состав статьи необходимо включать:

– УДК (см., например, УДК по электронному адресу: http://www.naukapro.ru/metod.htm);

– фамилии и инициалы авторов на русском и английском языках;

– название на русском и английском языках;

– аннотацию, как правило, объемом 200–270 слов на русском и английском языках;

– ключевые слова (5–7 слов или словосочетаний) на русском и английском языках;

– список литературы на русском языке и его транслитерацию латинской графикой. Список литературы необходимо оформлять в соответствии с требованиями ГОСТ Р 7.0.5–2008. Рекомендуемое число ссылок в одной статье: 15–20. Ссылки на работы, находящиеся в печати, не приводятся;

– сведения об авторах, включающие фамилию, имя, отчество, ученую степень, ученое звание (полностью), место работы с указанием почтового адреса, телефона организации и адреса электронной почты, должности, контактного телефона, сферу научных интересов и число опубликованных научных работ. Все эти данные помещаются на отдельной странице.

3. Статьи представляются в электронном варианте в виде файла формата *MS Word для Windows (*.doc)* по электронной почте на адреса rid@rosnou.ru и universitas@mail.ru. Название файла должно состоять из фамилии автора и названия статьи.

В тексте допускаются выделения шрифтами: **полужирный прямой**, ***полужирный курсив***, *светлый курсив*. Примеры рекомендуется выделять *курсивом*; заголовки, подзаголовки, новые термины и понятия – полужирным шрифтом.

Не рекомендуется использовать для выделения элементов текста ПРОПИСНЫЕ БУКВЫ, р а з р я д к у через пробел и подчеркивание, а также подстрочные ссылки.