Методология

КОЛИЧЕСТВЕННЫЕ МЕТРИКИ ДЛЯ ОЦЕНКИ КАЧЕСТВА КВАНТОВАНИЯ УЧЕБНОЙ ИНФОРМАЦИИ

Александр Рыбанов,
Волжский политехнический интитут (филиал) ФБГОУ ВПО «Волгоградский государственный технический университет»

В статье рассматриваются вопросы количественной оценки квантованного представления учебной информации на основе ряда метрик: сложность квантованного текста относительно исходного текста, коэффициент степени сжатия квантованного текста, релевантность квантованного текста исходному тексту. Процесс получения значений количественных метрик для квантованного представления учебной информации ориентирован на использование инструментальных средств: сжатия данных, автоматического реферирования, семантического анализа текста.
Сравнительный анализ квантованных текстов предлагается проводить с использованием методов равномерной оптимизации и справедливого компромисса.

Ключевые слова: квантование учебных текстов, учебный контент, степень сжатия, сложность квантованного текста, релевантность квантованного текста, автоматизация квантования, квазиреферат

Введение

Инструментальные средства разработки учебного контента отстают в развитии от систем дистанционного обучения (СДО). В свою очередь, успешность применения СДО зависит от качества и эффективной организации учебного контента.

Существующие в настоящее время СДО *Moodle*, *Ilias*, *Claroline*, *A-tutor* и др. не предоставляют разработчикам дистанционных учебных курсов возможность оценки качества учебного контента. Между тем, оценка учебного контента направлена на выявление достоинств и недостатков учебной информации и на принятие решения о необходимости и оптимальных условиях его использования в процессе дистанционного обучения.

Постановка задачи

Качество и эффективная организация учебного контента непосредственно влияют на следующие показатели[1] СДО:

1) *Коэффициент усвоения учебного контента* (K) — представляет собой отношение учебного контента, усвоенного пользователями СДО в течение определённой единицы времени, к контенту, сообщённому пользователям в течение этой единицы времени:

$$K = \frac{I_\theta}{I_\alpha},$$

где I_θ — усвоенный контент; I_α — сообщённый контент.

Если один и тот же контент усваивался пользователями в течение различных единиц времени, то коэффициент K следует разделить на время t. Для измерения I_α и I_θ можно использовать сравнительный анализ тезауруса пользователя СДО и тезауруса учебного контента[2].

2) *Скорость усвоения учебного контента* или соотношение коэффициента усвоения со временем усвоения:

$$K_i = \frac{t_i}{t_{cp}},$$

где K_i — коэффициент относительного учебного времени; t_i — время, затрачиваемое i-м пользователем СДО на усвоение определённого учебного контента; t_{cp} — среднее время усвоения определённого учебного контента группой пользователей СДО.

3) *Прочность усвоения учебного контента* — показывает уровень знаний, умений и навыков пользователя СДО по истечении некоторого времени после прохождения дистанционного курса:

$$\alpha_t = \frac{I_m}{I_\alpha},$$

где I_α — сообщённый контент; I_m — оставшийся в памяти и

1 *Рыбанов А.А.* Оценка качества текстов электронных средств обучения // Школьные технологии. 2011. № 6. С. 172–174.

2 *Рыбанов А.А.* Степень соответствия между тезаурусом учащегося и тезаурусом учебного контента как метрика процесса усвоения дистанционного учебного курса // Педагогические измерения. 2013. № 3. С. 77–91.

эффективно используемый пользователем по истечении некоторого времени t учебный контент.

Разработка учебного контента СДО включает в себя развитие технологий проектирования контента, таких как квантование[3] учебной информации. Коэффициенты K, K_i и α_m зависят, в том числе, и от качества квантования учебной информации[4].

Актуальной задачей является формирование системы количественных критериев для оценки качества квантования учебной информации.

Понятие процесса квантования

Квантование — это разделение учебной информации на элементарные фрагменты (учебные единицы, шаги, кадры) различного назначения (информационные, тренирующие, контролирующие, управляющие). Объём текстовой информации в этих фрагментах должен быть ограничен.

Процесс квантования — это преобразование

$$U' = f(U),$$

где $U = (u_i \mid i = \overline{1,n})$ — учебная информация, предназначенная для квантования, u_i — логически законченный фрагмент учебной информации U; $U' = (u'_i \mid i = 1,n)$ — квантованное представление учебной информации, где u'_i — квант учебной информации.

Принцип системного квантования учебной информации предполагает учёт следующих *закономерностей*:

- учебная информация большого объёма запоминается с трудом;
- учебная информация, представленная компактно, в определённой системе, лучше воспринимается;
- выделение в учебной информации смысловых единиц способствует эффективному запоминанию.

Учитывая то, что квант u'_i учебной информации должен содержать наиболее информативную часть фрагмента u_i, требования к кванту учебной информации можно формализовать следующим образом:

- квант u'_i учебной информации должен обладать более низкой избыточностью и большей энтропией, чем u_i;
- квант u'_i учебной информации по объёму должен быть меньше соответствующего ему фрагмента u_i учебной информации: $|u'_{i|}| \leq |u_i|$.

Неавтоматизированный процесс построения педагогом кванта u'_i для фрагмента u_i учебной информации состоит из следующих этапов:

3

Аванесов В.С. Применение заданий в тестовой форме и квантованных учебных текстов в новых образовательных технологиях // Педагогические измерения. 2012. № 2. С. 75–91.

4

Рыбанов А.А. Алгоритмическое и математическое обеспечение автоматизированной системы оценки качества учебного процесса по контрольным картам // Вестник компьютерных и информационных технологий. 2009. № 2. С. 30–36.

• подготовительный: чтение и осмысление фрагмента u_i учебной информации;
• аналитический: выделение основных смысловых единиц (предложения, слова, словосочетания), построение структуры кванта u_i' для фрагмента u_i учебной информации);
• непосредственное построение кванта u_i для фрагмента u_i учебной информации (выделенные ранее единицы располагаются в единый вторичный текст в соответствии со структурой кванта u_i').

В качестве смысловых единиц кванта u_i' для фрагмента u_i учебной информации могут быть:
• γ_1: полное (без изменений) ключевое предложение исходного текста ;
• γ_2: перефразированное ключевое предложение исходного текста u_i;
• γ_3: предложение из ключевых слов и словосочетаний исходного текста u_i;
• γ_4: предложение, обобщающее несколько предложений исходного текста u_i .

Автоматизация процесса квантования учебной информации

Автоматизация процесса квантования учебной информации возможна на основе применения средств автоматического реферирования[5].

Автоматическое реферирование (automatic text summarization) — это составление коротких изложений материалов. Подходы к решению данной задачи можно разделить на две группы: *квазиреферирование* и краткое *изложение содержания первичных документов*.

Квазиреферирование основано на экстрагировании фрагментов документов — выделении наиболее информативных фраз и формировании из них квазирефератов.

Краткое изложение исходного материала[6] основывается на выделении из текстов с помощью методов искусственного интеллекта и специальных информационных языков наиболее важной информации и порождении новых текстов, содержательно обобщающих первичные документы.

Рассмотрим квазиреферирование, как один из возможных подходов для автоматизации процесса квантования учебной информации. Данному подходу свойственно выделение смысловых единиц γ_1 и γ_3.

При квазиреферировании общий вес текстового блока определяется по формуле:

$$Weight = Location + KeyPhrase + StatTerm,$$

где *Location* — коэффициент, который определяется расположением блока в исходном

5 *Мишуков А.А.* Обзор систем автореферирования общего профиля //Информационное противодействие угрозам терроризма. 2005. № 4. С. 34–38.

6 *Герте Н.А., Курушин Д.С., Нестерова Н.М.* Свёртывание информации в процессе реферирования: методы и возможные пути формализации // Вестник Пермского национального исследовательского политехнического университета. Проблемы языкознания и педагогики. 2013. № 7 (49). С. 188–196.

тексте; *KeyPhrase*— весовой коэффициент ключевой фразы, представляющей собой конструкции-маркеры, которые резюмируют, типа «в заключение», «в данной статье», «в результате анализа» и т.п.; *StatTerm* — статистический вес текстового блока, вычисляемый как нормированная по длине блока сумма весов входящих в него строк (слов и словосочетаний).

В основе квазиреферирования лежат методы, использующие для выделения наиболее значимых предложений *деревья решений*[7], *скрытые марковские модели*[8], *логлинейные модели*[9], *нейронные сети*[10].

Метрики качества квантования учебной информации

Квантованные тексты U' должны обладать меньшей избыточностью и большей энтропией по сравнению с исходными U. Поэтому в качестве метрик качества квантования учебной информации предлагается использовать:

- сложность квантованного текста U' относительно текста U;
- коэффициент степени сжатия квантованного текста U';
- релевантность квантованного текста U' исходному тексту U.

Информационное определение энтропии через сложность ввёл А.Н. Колмогоров[11]: Сложностью последовательности букв A является длина (в двоичном алфавите) минимальной программы, которая выводит A, а энтропия A — это её сложность, деленная на длину в битах. Сложность текста по Колмогорову можно вычислить, воспользовавшись программами сжатия данных, например компрессором 7-zip 9.20 (www.7-zip.org). Данный компрессор реализует алгоритм сжатия LZMA (*Lempel-Ziv-Markov chain-Algorithm*), который относится к словарным алгоритмам сжатия информации без потерь, базирующимся на алгоритме Лемпеля–Зива.

Относительную сложность текста U' относительно текста U определим следующим образом:

1) сожмём текст U и измерим длину получившегося архива $C(U)$;

2) сожмём текст $U + U'$, получившийся присоединением текста U к тексту U, и измерим длину $C(U + U')$получившегося архива;

3) относительную сложность текста U' относительно текста U определим как $C(U'|U) = C(U + U') - C(U)$.

Чем меньше величина $C(U'|U)$, тем больше текст U' зависит от текста U.

Степень избыточности учебной информации можно

7
Lin C.-Y.
Training a selection function for extraction // Proceedings of CIKM '99. 1999. С. 55–62.

8
Conroy J.M., O'leary D.P.
Text summarization via hidden markov models // Proceedings of SIGIR '01. 2001. С. 406–407.

9
Osborne M.
Using maximum entropy for sentence extraction // Proceedings of the ACL'02 Workshop on Automatic Summarization. С. 1–8.

10
Svore K., Vanderwende L., Burges C.
Enhancing single-document summarization by combining RankNet and third-party sources // Proceedings of the EMNLP–CoNLL. 2007. С. 448–457.

11
Колмогоров А.Н.
Три подхода к определению понятия «количество информации» // Проблемы передачи информации. 1965. Т. 1. №1б. С. 3–11.

оценить через степень сжатия файла (текста), которая характеризуется коэффициентом K_c, определяемым как отношение объёма сжатого файла V_c к объёму исходного файла V, выраженное в процентах:

$$K_c = \frac{V_c}{V} \cdot 100\%.$$

В качестве меры схожести двух текстов U и U воспользуемся релевантностью, рассчитываемой по формуле:

$$R(U',U) = \sum_{j=1}^{n} w_j \cdot w'_j,$$

где j — номер ключевого слова в исходном тексте U; w_j — вес j-го ключевого слова в тексте U; w'_j — вес j-го ключевого слова в квантованном тексте U'.

Проведем анализ качества квантования учебной информации на примере статьи Ирины Веренчик[12]. В табл. 1 приведены статистические показатели для произведения А.П. Чехова «Белолобый». Текст данного произведения был разделён на семь фрагментов.

Полученные фрагменты учебной информации U были подвергнуты неавтоматизированному процессу квантования, выполненному педагогом. Метрические характеристики для полученных квантов учебной информации U' приведены в табл. 2.

Для автоматизированного процесса квантования был использован онлайн-сервис referat.keywordrush.com (рис. 1), предназначенный для построения квазиреферата. Метрические характеристики для полученных квантов учебной информации U'' приведены в табл. 3.

Для расчёта значений релевантности $R(U',U)$ и $R(U'',U)$ применялся онлайн-сервис для семантического анализа текста: seozor.ru/tools/analyzer.php. Данный сервис позволяет определить необходимое для расчёта релевантности множество

12

Веренчик И.И. Квантование текста и разработка заданий в тестовой форме. На примере произведения А.П. Чехова «Белолобый»// Педагогические измерения. №1, 2012. С. 98–105.

Таблица 1

Метрики для исходной учебной информации *U*

Фрагмент учебной информации	Слов	Символов	Абзацев	Предложений	Предложений в абзаце	Слов в предложении	Символов в слове	V_c, байт	V, байт	K_c, %
Старая волчиха	189	1115	3	8	2,6	23,6	4,7	729	1119	65,15
Зимовье Игната	207	1236	3	16	5,3	12,9	4,7	807	1240	65,08
Переполох в хлеву	139	872	4	9	2,2	15,4	5,0	598	878	68,11
Ненужная добыча	170	994	1	8	8,0	21,2	4,6	655	994	65,90
В волчьем логове	277	1673	6	13	2,1	21,3	4,8	1055	1683	62,69
Щенок и волчата	268	1609	7	18	2,5	14,8	4,7	993	1624	61,15
Возвращение домой	396	2427	13	36	2,7	11,0	4,8	1414	2451	57,69

Таблица 2

Метрики для квантованного представления *U′* учебной информации *U*

Фрагмент учебной информации	Метрики для квантованного представления учебной информации V'			$\frac{V'}{V}$, %	$C(U'\|U)$, байт
	V'_c, байт	V', байт	K'_c, %		
Старая волчиха	284	362	78,45	32,35	18
Зимовье Игната	494	691	71,49	55,73	17
Переполох в хлеву	295	398	74,12	45,33	13
Ненужная добыча	334	451	74,06	45,37	13
В волчьем логове	468	635	73,70	37,73	31
Щенок и волчата	550	810	67,90	49,88	43
Возвращение домой	893	1443	61,88	58,87	40

Рис. 1. Онлайн-сервис автоматического реферирования (referat.keywordrush.com)

ключевых слов и их весов по заданному фрагменту текста. Пример расчёта релевантности для фрагмента текста *Старая волчиха* приведен в табл. 4.

Значения релевантности $R(U',U)$ и $R(U'',U)$ для квантованных текстов приведены в табл. 5.

Таблица 3

Метрики для квантованного представления *U˝* учебной информации *U*

Фрагмент учебной информации	Метрики для квантованного представления учебной информации V''			$\frac{V''}{V}$, %	$C(U''\|U)$, байт
	V''_c, байт	V'', байт	K''_c, %		
Старая волчиха	337	456	73,90	40,75	17
Зимовье Игната	525	778	67,48	62,74	21
Переполох в хлеву	374	533	70,17	60,71	28
Ненужная добыча	486	711	68,35	71,53	16
В волчьем логове	512	754	67,90	44,80	38
Щенок и волчата	645	1043	61,84	64,22	34
Возвращение домой	1142	1892	60,36	77,19	44

Таблица 4

Расчёт меры схожести квантованных текстов для фрагмента *Старая волчиха*

j	Ключевое слово	w_j	w'_j	w''_j	$w_j \cdot w'_j$	$w_j \cdot w''_j$
1	ВОЛЧОНОК	1,8	0	0	0	0
2	ДЕРЕВО	1,8	0	0	0	0
3	ДОРОГА	1,8	0	4,26	0	7,67
4	ДРУГ	1,8	5	4,26	9	7,67
5	ВОЛЧИХА	1,8	5	0	9	0
6	СЛЕД	1,8	0	4,26	0	7,67
7	ЗДОРОВЬЕ	1,8	0	0	0	0
Мера схожести *R*					18	23,01

Таблица 5

Меры схожести *R*(*U´*, *U*) и *R*(*U˝*, *U*)

Фрагмент учебной информации	$R(U',U)$	$R(U'',U)$
Старая волчиха	23,01	18,00
Зимовье Игната	58,72	41,77
Переполох в хлеву	72,68	55,5
Ненужная добыча	62,38	33,90
В волчьем логове	37,05	31,05
Щенок и волчата	60,46	51,07
Возвращение домой	37,86	43,71

Сравнительный анализ результатов неавтоматизируемого и автоматизируемого процесса квантования

Критерии K_c, C и R имеют различные масштабы и шкалы измерения, поэтому, прежде чем приступить к решению многокритериальной задачи, их необходимо привести к одной единице измерения. Предлагается следующий способ получения безразмерной формы критериев:

$$f_j^H(A_i) = \frac{f_j(A_i) - \min\limits_i \{f_j(A_i)\}}{\max\limits_i \{f_j(A_i)\} - \min\limits_i \{f_j(A_i)\}},$$

где $j = \overline{1,n},\ \min\limits_i \{f_j(A_i)\} \neq \max\limits_i \{f_j(A_i)\}.$

Так как критерий C минимизируется, то для того, чтобы все критерии стремились к максимуму, умножим безразмерные величины критерия C на (–1), и добавим к нему константу, например 1. Значения нормированных критериев K_c, C и R приведены в табл. 6.

Сравнительную оценку качества квантованных представлений U' и U'' учебной информации U выполним с применением методов многокритериального выбора. Интегральный критерий для равнозначных критериев K_c, C и R рассчитаем по методу равномерной оптимизации и методу справедливого компромисса.

Интегральный критерий выбора по методу равномерной оптимизации вычисляется по формуле:

$$f(A^*) = \max_i \{\sum_{j=1}^{n} f_j^H(A_i)\}.$$

Интегральный критерий выбора по методу справедливого компромисса вычисляется по формуле:

$$f(A^*) = \max_i \{\prod_{j=1}^{n} f_j^H(A_i)\}.$$

Анализ значений интегральных критериев показывает, что U' является лучшим, по сравнению с U'', квантованным представлением для учебной информации U.

Заключение

Проведённый анализ метрических характеристик квантованного представления U' для учебной информации U позволяет сделать следующие выводы:

1) чем выше величина $R(U', U)$, тем больше U' соответствует U.

2) чем меньше величина $C(U'|U)$, тем больше текст U' зависит от текста U.

3) чем выше степень сжатия K_C' для U', тем меньше избыточность текста U', следовательно для U' сжатия всегда больше чем для U.

Таблица 6

Сравнительный анализ квантованных представлений *U′* и *U′′* учебной информации *U*

Фрагмент учебной информации	Квантованное представление учебной информации	$f_1^H(K_{ci})$	$f_2^H(C_i)$	$f_3^H(R_i)$	Метод равномерной оптимизации	Метод справедливого компромисса
Старая волчиха	*U′*	1,00	0,84	0,09	1,93	0,076
	U′′	0,75	0,87	0,00	1,62	0
Зимовье Игната	*U′*	0,62	0,87	0,74	2,23	0,399
	U′′	0,39	0,74	0,43	1,56	0,124
Переполох в хлеву	*U′*	0,76	1	1,00	2,76	0,760
	U′′	0,54	0,52	0,69	1,75	0,194
Ненужная добыча	*U′*	0,76	1	0,81	2,57	0,616
	U′′	0,44	0,9	0,29	1,63	0,115
В волчьем логове	*U′*	0,74	0,42	0,35	1,51	0,109
	U′′	0,42	0,19	0,24	0,85	0,019
Щенок и волчата	*U′*	0,42	0,03	0,78	1,23	0,010
	U′′	0,08	0,32	0,60	1	0,015
Возвращение домой	*U′*	0,08	0,13	0,36	0,57	0,004
	U′′	0,00	0	0,47	0,47	0

Предложенные в статье процедура получения метрик K_c, C, R и методика сравнительного анализа двух квантованных представлений для одного и того же учебного текста могут быть использованы при проектировании учебного контента систем дистанционного обучения.