УДК 004.912

*Н. С. Завражнев, А. С. Евдокимов, И. Р. Рудь**

ИЗВЛЕЧЕНИЕ СЕМАНТИЧЕСКИ ЗНАЧИМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ ИЗ АУДИО- И ВИДЕОФАЙЛОВ НА ОСНОВЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА С УЧЕТОМ НЕЧЕТКИХ КОЛЛОКАЦИЙ

В современном мире все чаще при решении различного рода производственных и управленческих задач сотрудники компаний сталкиваются с необходимостью анализа аудио- и видеоинформации (АВИ) больших объемов. Проблема анализа такой информации заключается в том, что для нормального ее восприятия невозможно сократить время обработки, в отличие от текстовой информации, где скорость чтения играет большую роль. При выполнении производственных и управленческих задач время является критическим ресурсом. Потому обработка больших объемов АВИ непосредственно сопряжена с финансовыми затратами и, зачастую, компаниям приходится нанимать дополнительный персонал, который необходим для качественной работы с большими объемами АВИ. Таким образом, ускорение обработки информации, заключенной в аудио- и видеофайлах (АВФ) позволит снизить издержки работы подобных компаний. Поэтому актуальной является задача автоматического построения саммари, содержащего семантически значимую часть текстовой информации, изложенной в АВФ.

Для решения данной задачи предлагается двухэтапный алгоритм (рис. 1), в котором на первом этапе происходит построение текста на основе аудиодорожки АВФ, а на втором из полученного текстового документа формируется саммари, в котором заключена семантика исходного АВФ.

Рис. 1. Схема работы алгоритма

* Работа выполнена под руководством канд. техн. наук, доцента кафедры «Информационные системы и защита информации», ФГБОУ ВО «ТГТУ» Д. В. Полякова.

Вместе с тем, есть ряд проблем, стоящих на пути автоматизации извлечения семантически значимой текстовой информации. Во-первых, отсутствие знаков препинания и разбиения на предложения текста, представленного в АВФ. Во-вторых, низкое качество современных моделей, анализирующих текстовые документы на основе семантики.

Для решения первой проблемы предлагается рассмотреть возможность анализа пауз в звуковой дорожке АВИ. На основе выявления, учета и анализа пауз предлагается расставить псевдознаки препинания, то есть знаки, не отражающие реальное разбиение на предложения, но позволяющие разбить полученную текстовую информацию на группу семантически слабо связных элементов.

Решением второй проблемы может стать использование нечетких коллокаций и антологий в модели латентно-семантического анализа. Для этого целесообразно воспользоваться обобщенной векторно-пространственной моделью текстовой коллекции [1]. Под коллокацией понимается коллективная локация термов, то есть группа термов, расположенных рядом друг с другом [2]. Классически термы в коллокации располагаются непосредственно друг за другом. Однако есть работы [3], в которых термы могут располагаться на различном расстоянии друг относительно друга. Причем это расстояние может быть задано как некоторым целым числом, так нечетким числом.

Коллокации, в которых расстояние между термами задано нечетким числом, получили название нечетких коллокаций [3]. Под расстоянием между термами понимается число других термов, расположенных в тексте между заданными, составляющими коллокацию.

Пусть D – множество текстовых документов, $D = \{d_1, d_2, \ldots, d_N\}$, $|D|=N$, где $|.|$ – мощность множества. А T – множество термов, встречающихся в документах D. То есть будем считать, что если некоторый терм t встречается в документе $d_i, i = \overline{1,N}$, то $t \in T$. Пусть $T = \{t_1, t_2, \ldots, t_n\}$, $|T| = n$. Также сам факт появления терма t в d_i, будем обозначать $t \in d_i$, что, как может показаться на первый взгляд, не является тривиальным использованием символа $\in$, так как d_i нельзя назвать множеством термов. Действительно, в документе термы присутствуют в определенной последовательности, связанные синтаксически и морфологически.

Пусть в результате некоторого преобразования (1) документ d_i представлен в виде множества характеристических объектов (*p*). Так как в этом случае d_i потерял часть семантической информации, будем обозначать рассматриваемое множество $\hat{d}_i$, которое представляет собой некоторую оценку d_i. Тогда, для удобства работы с характеристическими объектами, введем следующее обозначение:

$$\hat{d}_i = \left\{ p_1^i, p_2^i, ..., p_{M_i}^i \right\}, \quad (1)$$

где $\left|\hat{d}_i\right| = M_i$. Элементы множества $\hat{d}_i$ (характеристические объекты) представляют собой встреченные в документе артефакты (конструкции, структуры), в большей или меньшей степени характеризующие семантику d_i.

Причем, если соответствующие артефакты входят в состав документа неоднократно, то и в (2) они появляются соответствующее число раз. Представление документа в виде (2) задает универсальное множество характеристических объектов $U_p = \bigcup_{i=1}^{N} \hat{d}_i$. Действительно, если первоначально определить универсум U_p, то множество $\hat{d}_i, i = \overline{1, N}$ определяется как

$$\hat{d}_i = \left\{ p \in U_p \mid p \in d_i \right\}, \quad (2)$$

причем количество вхождений объекта *p* в $\hat{d}_i$ равно числу его появлений в документе d_i.

Рассмотрим множество $U_F = \left\{ \tilde{F}_1, \tilde{F}_2, ..., \tilde{F}_K \right\}$, $\left|U_F = K\right|$. U_F представляет собой универсальное множество характеристик (пространство факторов), по которым оценивается семантический смысл текстового документа. Каждый элемент U_F – нечеткое множество $\tilde{F}_i, i = \overline{1, K}$, задаваемое функцией принадлежности $\mu_i : U_p \to [0,\ 1]$, то есть $\tilde{F}_i \subset U_p, \forall i = \overline{1, K}.$

В [1] была предложена и обоснована обобщенная векторно-пространственная модель текстовой коллекции в виде матрицы, каждый элемент которой имеет вид:

$$f_j^i = T\left(\xi\left(\frac{\sum_{k=1}^{M_j}\mu_i\left(p_k^j\right)}{M_j}\right),\zeta\left(\frac{N}{1+\sum_{i=1}^{N}S\left(\mu_j\left(p_1^i\right),\mu_j\left(p_2^i\right),\ldots,\mu_j\left(p_{M_i}^i\right)\right)}\right)\right), \qquad (3)$$

где ξ и ζ – произвольные функции, такие что ξ: [0, 1] $\rightarrow$ [0, 1], $\xi(0)=0$, $\xi(1)=1$, ζ: [1, |D|] $\rightarrow$ [0, 1], $\zeta(1)=0$, $\zeta(|D|)=1$ и $\zeta\uparrow$ на [1, |D|].

Заметим, что при конкретных значениях функций (3) следующего вида: $\xi(x)=x, \forall x\in R$, $\zeta(x)=\log(x), \forall x\in R$, $T(x,y)=xy$, $\forall x,y\in R$, а $S(x,y)=x+y-xy$, выражение (3) целесообразно переписать как

$$f_j^i = \frac{\sum_{k=1}^{M_j}\mu_i\left(p_k^j\right)}{M_j}\log\left(\frac{N}{1+\sum_{i=1}^{N}S\left(\mu_j\left(p_1^i\right),\mu_j\left(p_2^i\right),\ldots,\mu_j\left(p_{M_i}^i\right)\right)}\right), \qquad (4)$$

Для (4) было показано, что при $U_p = T$ представление (1) в силу (2) совпадает непосредственно с текстом, а в качестве артефактов выступают только термы. Причем полученный частный случай (4) является известной векторно-пространственной моделью текстовой коллекции.

То есть обобщенная модель (4) позволяет проводить латентно-семантический анализ на основе не только термов, но и любых других артефактов, в том числе нечетких коллокаций и антологий.

В рамках дальнейших исследований предполагается провести научно-исследовательские работы по созданию моделей, методов и алгоритмов построения и оценки семантической значимости нечетких коллокаций, фаззификации антологий, а также опытно-конструкторские работы по разработке программного обеспечения в соответствии со схемой, представленной на рис. 1.

Разработанное программное обеспечение будет предоставлять сервис построения саммари (короткого пересказа) аудио информации,

представленной на АВФ. Для решения большинства задач, представленных на рис. 1, планируется найти готовые решения, в том числе основанные на технологиях машинного обучения. Вместе с тем, для формализации трех последних блоков планируется разработать программные модули, работающие на основе модели (4), а также использующие результаты запланированных исследований нечетких коллокаций.

Список литературы

1. Поляков, Д. В. Обобщение векторно-пространственной модели для оценки семантической значимости характеристик текстовых документов / Д. В. Поляков, Н. М. Митрофанов, Е. Н. Лепешкин // Приборы и системы. Управление, контроль, диагностика. – 2016. – № 2 – С. 51 – 61.

2. Ягунова, Е. В. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов / Е. В. Ягунова, Л. М. Пивоварова // Сб. НТИ, Сер. 2, № 6. М., 2010. – URL : http://webground.su/services.php? param=priroda_collac &part=priroda_collac.htm.

3. Поляков, Д. В. Метод формализации нечетких коллокаций термов в текстах на основе лингвистических переменных / Д. В. Поляков, Н. М Митрофанов, А.С. Матвеева. // Прикаспийский журнал: Управление и высокие технологии. – 2015. – № 4(32) – С. 167 – 183.

Кафедра «Информационные системы и защита информации»
ФГБОУ ВО «ТГТУ»