УДК 004.5:004.91

А. Е. Риман, асп.;
С. Н. Сердюк, к. т. н., доц.

СИСТЕМА РЕФЕРИРОВАНИЯ МУЛЬТИМОДАЛЬНОЙ ИНФОРМАЦИИ

Рассмотрен процесс реферирования мультимодальной информации. То есть исходная информация может представляться в виде текста, изображения, аудио или видео. Предлагается модель процесса реферирования, улучшенная за счёт введения дополнительного этапа, позволяющего обрабатывать мультимодальную информацию. Приводятся алгоритм функционирования системы реферирования на этом этапе и модель процесса преобразования исходного документа во внутренний формат системы.

Введение

В последнее время, благодаря развитию систем документооборота, Internet и ряда других факторов, наблюдается накопление огромных массивов неформализованных документов. Однако быстро получить требуемую информацию, необходимую руководству организации, достаточно сложно вследствие того, что информационные потоки внутри организации спланированы заранее, сроки представления отчетности регламентированы. В итоге, возрастает количество нестандартных отчетов, которые, по требованию руководства, готовятся на заказ системным аналитиком организации [1]. Одним из методов, призванных помочь в поиске и обработке необходимой информации, является автоматическое реферирование.

Реферирование, или составление аннотаций, — это процесс извлечения наиболее важной информации из одного или нескольких источников для составления их сокращенной версии для потребностей определенных пользователей или задач [2].

История применения вычислительной техники для реферирования насчитывает свыше 50 лет и связана с именами таких исследователей, как Г. П. Лун [3], Г. Эдмунсон [4], В. Е. Берзон, И. П. Севбо, Э. Ф. Скороходько, Д. Г. Лахути, Р. Г. Пиотровский и др. [5]. За эти годы выработаны многочисленные подходы к решению данной проблемы [6—9]. Всплеск активности в данной области начался в конце 80-х годов XX века. Новый импульс дало создание и развитие сети Интернет, в которой далеко не каждый документ снабжен авторским резюме и ключевыми словами. Хотя некоторые производители уже сейчас предлагают инструменты для реферирования, объем информации в Сети растет и оперативно получать ее корректные сводки становится все сложнее [10]. Возможности предлагаемых инструментов ограничены выделением и выбором оригинальных фрагментов из исходного документа и соединением их в короткий текст. Подготовка же краткого изложения предполагает передачу основной мысли текста, и не обязательно теми же словами. Поэтому проблема связного выбора и выражения информации в автоматическом реферировании остается актуальной и на сегодняшний день [11].

Упомянутые инструменты реферирования рассчитаны на обработку только текстовой информации. Однако источники информации далеко не всегда являются текстами [10]. Ведь необходимо подготавливать аннотации и на видеозаписи, к примеру, спортивных соревнований, и на аудиозаписи, например, формировать дайджест новостей, переданных по радио. Исходной информацией может быть и изображение, например, график [12] или отсканированный документ. Таким образом, необходима система, которая бы могла обрабатывать информацию различной модальности — текст, изображения, аудио или видео.

Постановка задачи

Широко распространенная модель процесса реферирования описывается тремя этапами [13]:

1) формализация исходного текста;
2) преобразование формальной модели текста в формальную модель реферата;
3) генерация текста реферата из его формальной модели.

Однако эта модель не позволяет обрабатывать мультимодальную информацию, а рассчитана на работу только с текстовой информацией. Чтобы устранить указанный недостаток, предлагается улучшить данную модель введением начального (0) этапа — перевод исходного документа в текстовый формат, пригодный для системы автоматического реферирования текстов.

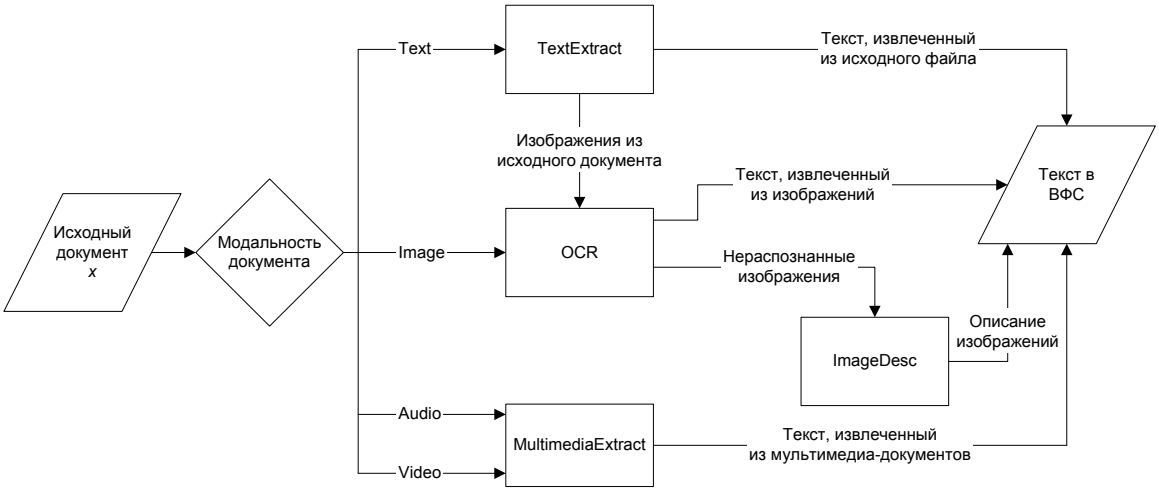
Целью данной работы является разработка алгоритма функционирования системы реферирования мультимодальной информации на этапе 0 и модели процесса преобразования исходного документа во внутренний формат системы.

Метод решения задачи

Рассмотрим процесс преобразования исходной мультимодальной информации в текстовый формат, пригодный для системы автоматического реферирования текстов — внутренний формат системы (ВФС).

ВФС должен обеспечивать минимальные потери значимой информации. Этой информацией может быть форматирование текста (шрифт, размер, выделение), структура (абзацы, разделы, страницы), различные уровни заголовков, ключевые слова, метаданные и др. Вся эта информация может быть использована на последующих этапах процесса реферирования. Например, при использовании методов реферирования, учитывающих статистическую важность слов текста [14], указанные выше параметры и элементы исходного документа могут играть существенную роль в выборе необходимых предложений для составления реферата. Также при преобразовании документа в ВФС необходимо учитывать содержащиеся в нём изображения. Так, согласно исследованию [15], до 50% информации, содержащейся в статьях по биологии, связана с изображениями (рисунками, графиками, диаграммами и т.п.).

Процесс преобразования исходного документа в ВФС изображен на рисунке.



Процесс преобразования исходного документа в ВФС

Пусть исходная информация представляется в виде некоторого файла (документа) $x \in X$, где X — множество файлов.

Обозначим через F — множество форматов файлов. Тогда функция

$$\Phi : X \to F$$

позволяет определить формат исходного файла

$$f = \Phi(x); \quad f \in F.$$

Введем $M = \{Text, Image, Audio, Video\}$ — множество модальностей исходных файлов.

Функция *FileMod* определяет модальность исходного файла по его формату:

$$FileMod : F \to M$$

$$m = FileMod(f); \quad m \in M$$

В таблице приведено соответствие возможных форматов документа $f_i \in F$ и модальностей $m_i \in M$.

Соответствие форматов документов и модальностей

Формат документа f_i^*	Модальность документа m_i	Описание модальности
Plain Text, HTML, XML, RTF, DOC, PDF, PS…	Text	Текстовый документ
BMP, GIF, JPEG, TIFF, PNG…	Image	Изображение
WAVE, MP3, WMA, AAC, OGG…	Audio	Аудио-файл
MPEG, AVI, ASF, QT, RM, WMV…	Video	Видео-файл

В зависимости от модальности документа, к исходному файлу *x* применяется соответствующая ей функция преобразования в ВФС.

Определим текст в ВФС как подмножество t_i множества *T* различных текстовых элементов и их параметров:

$$t_i \subset T;\ \ t_i \in P(T),$$

где *P(T)* — множество всех подмножеств множества *T*.

При *m = Text* к *x* применяется функция *TextExtract*:

$$te = TextExtract(x);\quad te = \langle t_1, im_1 \rangle,$$

где $t_1 \in P(T)$ — текст в ВФС, извлеченный из исходного документа; $im_1 \in P(I)$ — множество изображений, извлеченных из исходного файла, *I* — множество различных изображений.

При этом в t_1 вводятся элементы $ti_j \in T$, которые при дальнейшей обработке будут заменены на текст, определяемый изображениями из im_1.

Наличие изображений обусловлено тем, что большинство текстовых документов может содержать их, например, в виде иллюстраций к основному тексту, или сам текст может храниться в виде растрового изображения.

При *m = Image* исходный документ *x* представляет собой подмножество из *I*, т. е. в данном случае $x \in P(I)$. В этом случае, а также при $im_1 \neq \varnothing$ множество изображений обрабатывается функцией *OCR*, которая производит распознавание текста во входных изображениях:

$$OCR:\ P(I) \to P(T) \times P(I),$$

$$\langle t_2, im_2 \rangle = OCR(x),$$

$$\langle t_3, im_3 \rangle = OCR(im_1),$$

где t_2, $t_3 \in P(T)$ — текст, извлеченный из входных изображений; im_2, $im_3 \in P(I)$ — множества нераспознанных изображений, т. е. входные изображения или их части, которые были определены этой функцией как «не текстовые».

Если множество im_2 (im_3) не пусто, то к нему применяется функция *ImageDesc*, которая производит классификацию изображений и выделяет из них различные ключевые слова или составляет их описание:

$$ImageDesc\ :\ P(I) \to P(T);$$

$$\langle t_{4,5} \rangle = ImageDesc(im2,3),$$

где $t_{4,5} \in P(T)$ — множество ключевых слов, определенных по входным изображениям.

Если $m \in \{Audio, Video\}$, то к входному документу *x* применяется функция преобразования мультимедийных данных в текстовый вид

$$MultimediaExtract: X \to P(T),$$

$$MultimediaExtract(x) = \langle t_6 \rangle,$$

где $t_6 \in P(T)$ — текст, извлеченный из входных мультимедиа-документов.

Таким образом, в результате применения указанных выше функций преобразования исходного документа в текст в ВФС, множества t_i, $i = 1 \div 6$ объединяются в одно множество $t_{\text{ВФС}} = \bigcup_{i=1}^{6} t_i$,

которое и является результатом работы системы реферирования мультимодальной информации на этапе 0.

Выводы

Приведена улучшенная модель процесса реферирования [13], позволяющая обрабатывать мультимодальную информацию. Разработаны алгоритм функционирования системы реферирования мультимодальной информации на этапе 0 и модель процесса преобразования исходного документа во внутренний формат системы. На их базе реализована программная система, позволяющая подключать сторонние модули, которые могут выполнять различные функции преобразования исходного документа в ВФС.

СПИСОК ЛИТЕРАТУРЫ

1. Сердюк С. Н., Поздняков А. А. Анализ и синтез систем поддержки принятия решений // Радіоелектроніка. Інформатика. Управління. — 2000. — № 1. — С. 106—111.

2. Maybury, M. Generating Summaries from Event Data // Information Processing and Management. — 1995. — № 31(5). — P. 735—751.

3. Luhn, H. P. The automatic creation of literature abstracts // IBM Journal of Research and Development. — 1958. — № 2(2). — P. 159—165.

4. Edmundson, H. P. New methods in automatic abstracting // The Association for Computing Machinery. — 1969. — № 16(2).

5. Блюменау Д. И., Афанасова Л. Н. Индикаторный метод компьютерного свертывания в процессе обучения аналитико-синтетической передачи информации // Научные и технические библиотеки. — 2001. — № 12. — С. 29—42.

6. Paice, C.D. Constructing literature abstracts by computer: techniques and prospects // Information Processing and Management. — 1990. — № 26(2). — P. 171—186.

7. Endres-Niggemeyer, B.; Hobbs, J.; and Sparck Jones, K. eds. Summarising text for intelligent communication. Dagstuhl Seminar Report 79, 13.12-17.12.93 (9350), IBFI, Schloss Dagstuhl, Wadern, Germany, 1995.

8. Sparck Jones, K. and Endres-Niggemeyer B. Special Issue: Summarizing Text // Information Processing and Management. — 1995. — № 31(5). — P. 625—784.

9. Mani, I. and Maybury, M. eds. 1997. Intelligent scaleable text summarization // Proceedings of a Workshop Sponsored by the ACL. Somerset NJ: Association for Computational Linguistics.

10. Hahn, U. and Mani, I. The challenges of automatic summarization // IEEE Computer. — 2000. — № 33(11). — P. 29—36.

11. F. Johnson. Automatic abstracting research // Library Review. — 1995. — № 44(8). — P. 28—36.

12. Carberry, S.; Elzer, S.; Green, N.; McCoy, K. and Chester, D. Extending Document Summarization to Information Graphics // Proceedings of the ACL Workshop on Text Summarization. — 2004. — P. 3—9.

13. Sparck-Jones, K. Automatic summarizing: factors and directions // Advances in Automatic Text Summarization / Edited by Mani, I. and Maybury, M. — MIT Press. — 1999. — P. 1—12.

14. Fresno, V.; Ribeiro, A. An Analytical Approach to Concept Extraction in HTML Environments // Journal of Intelligent Information Systems. — 2004. — № 22(3). — P. 215—235.

15. Futrelle, R. P. Handling Figures in Document Summarization // Proceedings of the ACL Workshop on Text Summarization. — 2004. — P. 61—65.

Риман Александр Ефимович — аспирант; ***Сердюк Сергей Никитович*** — доцент.

Кафедра программных средств, Запорожский национальный технический университет