М388331

Министерство высшего и среднего специального образования СССР

Московское
ордена Ленина, ордена Октябрьской Революции и ордена Трудового
Красного Знамени высшего технического училища имени Н.Э.Баумана

НА ПРАВАХ РУКОПИСИ НЕ ВЫДАЕТСЯ

На правах рукописи

ШАТОХИНА Светлана Николаевна

ИССЛЕДОВАНИЕ И РАЗРАБОТКА ПРИНЦИПОВ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ АУДИОВИЗУАЛЬНОЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ ЭВМ

/ 05.13.06 - Автоматизированные системы переработки информации и управления. /

А в т о р е ф е р а т

диссертации на соискание ученой степени
кандидата технических наук

Москва 1981

Работа выполнена в Главном информационно-вычислительном центре Гостелерадио СССР.

Научный руководитель -
доктор технических наук, профессор СОЛОМАТИН Н.М.

Официальные оппоненты:
доктор технических наук, профессор Петров А.В.
кандидат технических наук, с.н.с. Енгалычев А.М.

Ведущая организация - Вычислительный центр Сибирского отделения АН СССР

Защита состоится 220282 на заседании специализированного совета К.053.15.04 МВТУ им. Н.Э.Баумана, ауд 504-10 в 16 час, по адресу: 107005, 2-я Бауманская 5

С диссертацией можно ознакомиться в библиотеке МВТУ им.Н.Э.Баумана

Автореферат разослан

Ученый секретарь совета
кандидат технических наук,

М 388331

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Современный этап научно-технической революции привел к необходимости оперативной переработки больших объемов аудиовизуальной информации /аудиовизуальных документов/*. Аудиовизуальные документы находят широкое применение в различных областях человеческой деятельности. Это рентгенограммы в медицине, аэрокосмическая видеоинформация - в метеорологии, геологии, океанографии и других науках о Земле, фонограммы - в биологии и космической связи, фотоснимки - в микробиологии, ядерной физике и т.д.

Особенно следует сказать о таком важном использовании аудиовизуальных документов, как телевидение и радиовещание. Информационная мощность фонда аудиовизуальных документов Гостелерадио СССР сегодня составляет 0,7 млрд. байтов, при этом, приблизительно каждые 10 лет объем фонда удваивается. Телевидение и радиовещание прочно вошли в повседневную жизнь и стали могучим средством идейного, политического и художественного воспитания человека. На XXVI съезде КПСС перед Советским телевидением и радиовещанием были поставлены задачи дальнейшего расширения объемов телевизионных и радиопередач, повышения их качества и совершенствования технической базы.

Все это делает задачу автоматизированной обработки аудиовизуальных документов одной из актуальных проблем сегодняшнего дня.

Цель работы. Целью настоящей работы является исследование и разработка принципов автоматизированной обработки аудиовизуальной информации с использованием ЭВМ. При этом рассматривались следующие

---

* - Принимая, что носителем "аудиовизуальной информации" всегда является "аудиовизуальный документ", в работе оба понятия использованы как синонимы.

НТБ МГТУ им.Н.Э.Баумана
1388231
МВТУ им. Н.Э. Баумана БИБЛИОТЕКА

основные задачи: классификация аудиовизуальных документов; разработка алгоритма семантического преобразования первоисточника; построение формализованной модели аудиовизуальных документов; разработка программного обеспечения; создание информационной базы и др.

Методы исследования. При разработке общей классификации документов с учетом форм представления семантической информации был использован математический аппарат теории множеств и формальной логики. Для построения формализованной модели описания аудиовизуальных документов использовались методы математической статистики, теория вероятностей и теория графов. При разработке языка описания аудиовизуальных документов был использован раздел математической лингвистики: "Теория формальных грамматик".

Научная новизна. Анализ состояния указанной проблемы по данным литературы показал, что имеющиеся результаты можно рассматривать как ее частные решения. В работе предложен комплексный подход к решению данной проблемы, который, в свою очередь, позволил:

- разработать научную методику автоматизированной обработки аудиовизуальной информации с использованием ЭВМ;
- определить взаимосвязь аудиовизуальных документов с различными формами представления семантической информации;
- сформулировать алгоритм семантического преобразования первоисточника;
- построить универсальную модель описания аудиовизуальных документов;
- разработать формализованный язык описания аудиовизуальных документов;
- провести исследования и дать рекомендации построения информационной базы;

– разработать программное обеспечение для обработки аудиовизуальных документов.

Практическая ценность. На основании выполненных исследований разработана инженерная методика автоматизированной обработки аудиовизуальных документов. Построены рабочие модели описания аудиовизуальных документов; созданы грамматические средства свертывания информации аудиовизуальных документов и программное обеспечение ее обработки; сформирована информационная база. По результатам эксплуатации систем определены плотности входных и выходных информационных потоков, получены объемно-временные и эксплуатационные характеристики, дана их оценка. Разработанная система включена в общий технологический процесс подготовки и выпуска теле- и радиопередач.

Реализация работы. На основании полученных результатов в ГИВЦ Гостелерадио СССР было разработано и внедрено две системы автоматизированной обработки аудиовизуальных документов телевидения и радиовещания.

Апробация работы. Содержание отдельных разделов и работы в целом было доложено:

- на Международном симпозиуме ГИ УI OIRT, Таллин, 1970 г.,
- на Ш научно-технической конференции ВНИИТРа, Москва, 1973 г.,
- на IУ научно-технической конференции ВНИИТРа, Москва, 1974г.,
- на Всесоюзном отраслевом совещании Гостелерадио СССР по вопросам автоматизации, Таллин, 1980 г.,
- на научно-техническом семинаре "Вопросы разработки и внедрения систем обработки данных на базе ЕС ЭВМ", Москва, МДНТП, 1980 г.,
- на научных семинарах кафедры "Электронно-вычислительные машины и системы" в МВТУ им. Н.Э.Баумана.

Публикации. По результатам выполненных исследований опубликовано семь работ.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав и заключения, изложенных на 172 страницах машинописного текста, 41 рисунка, 7 таблиц, списка цитированной литературы и 9 приложений.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В о в в е д е н и и отмечается актуальность работы, дается краткий обзор состояния вопроса по рассматриваемой проблеме, акцентируется внимание на ее специфике, формулируется цель и задачи диссертации, дается перечень вопросов выносимых на защиту, приводится аннотация диссертации по главам.

П е р в а я г л а в а диссертации посвящена исследованию информационных семантических систем и разработке классификации документов с учетом форм представления семантической информации. В этой же главе раскрывается многоплановость общей проблемы автоматизированной обработки аудиовизуальных документов. Сформулированы ее основные задачи.

В основу разработки классификации документов взято три следующих положения:

1 - Приняты три основные формы представления семантической информации: *Т* - текстовая, *А* - аудиовизуальная, *И* - изобразительная.

2 - Выделены две части семантической информации документа:
СЧД - содержательная часть документа,
БЧД - библиографическая часть документа.

3 - Определены формы представления семантической информации в частях документа:

СЧД $\rightarrow f(T, A, И)$;

БЧД $\rightarrow f(T)$;

Учитывая направленность исследований, вводится детализация аудиовизуальной формы представления семантической информации на два вида: $A_з$ и $A_б$

Считаем, что семантическая информация аудиовизуального документа имеет форму $A_з$, если эта информация представлена содержанием звукозаписей или радиопередач, и имеет форму $A_б$, если представлена содержанием фото-, теле-, кино- и т.п. материалов.

Множество основных форм представления семантической информации определим как множество $M^* = \{m_1^*, m_2^*, m_3^*, m_4\}$. Тогда содержательная часть аудиовизуальных документов может быть выражена через одноместные предикаты $B(x^*)$, определенные на этом множестве M^*

$$C(n_i^*) = \overline{7}B(m_1^*) \wedge \overline{7}B(m_2^*) \wedge \overline{7}B(m_3^*) \wedge B(m_4^*);$$

$$C(n_{i+1}^*) = \overline{7}B(m_1^*) \wedge \overline{7}B(m_2^*) \wedge B(m_3^*) \wedge \overline{7}B(m_4^*);$$

$$C(n_{i+2}^*) = \overline{7}B(m_1^*) \wedge \overline{7}B(m_2^*) \wedge B(m_3^*) \wedge B(m_4^*);$$

Подробное математическое описание классификации документов дано с использованием понятий теории множеств и правил исчисления предикатов. Исследуемые аудиовизуальные документы, в этом случае, определены как:

$$\mathcal{D}_i \in (\mathcal{D}_\alpha^{**} \cup \mathcal{D}_\beta^{**} \cup \mathcal{D}_\gamma^{**}) \quad \ldots\ldots \quad /2/$$

Теория построения семантических систем позволяет описать работу любой автоматизированной системы обработки информации как некоторую последовательность выполняемых функций:

$$S \Rightarrow F_1 \vec{\vee} F_2 \vec{\vee} F_3 \quad \ldots\ldots\ldots \quad /3/$$

где

S - общая функция автоматизированной системы обработки информации,

F_1 - преобразование исходной информации, подлежащей обработке в системе,

F_2 - процесс машинной обработки информации,

F_3 - преобразование результирующей информации,

$\Rightarrow$ - "следует"

$\vee$ - векторное логическое сложение.

Раскрывая каждую функцию системы следующим образом:

$$F_1 \Rightarrow \mathcal{I}_1 \vee \mathcal{P} \vee \mathcal{I}_2; \quad F_2 \Rightarrow F_2' + F_2''; \quad F_3 \Rightarrow R \vee R\mathcal{D};$$

определим S как

$$S \Rightarrow \mathcal{I}_1 \vee \mathcal{P} \vee \mathcal{I}_2 \vee (F_2' + F_2'') \vee R \vee R\mathcal{D};$$

Отсюда, общая функция семантической системы автоматизированной обработки аудиовизуальных документов может быть определена:

$$S_{sem} \Rightarrow \mathcal{I}_{1(A)} \vee \mathcal{P}_{sem} \vee \mathcal{I}_{2(T)} \vee (F_2' + F_2'') \vee R_{(T)} \vee R\mathcal{D}_{(A)} \quad /4/$$

где

$\mathcal{I}_{1(A)}$ - первичная семантическая информация, представленная в аудиовизуальной форме,

$\mathcal{P}_{sem}$ - семантическое преобразование информации, основными принципами которого являются:

- принцип семантической топологии,
- принцип адекватности,
- принцип передачи семантической информации.

$\mathcal{I}_{2(T)}$ - вторичная семантическая информация, представленная в текстовой форме,

$R_{(T)}$ - результирующая информация, представленная в текстовой

форме,

$RD_{(A)}$ – информация релевантных документов, представленная в аудиовизуальной форме.

Сравнительный анализ ряда семантических систем, общая функция которых соответствует выражению /4/, показал, что в каждой системе предлагается свой, отличный от других, метод преобразования аудиовизуальной информации первоисточника в текстовую.

Следовательно, чтобы построить семантическую систему автоматизированной обработки аудиовизуальных документов, прежде всего должна быть разработана единая методика преобразования первичного документа, приемлемая как для формы A_3, так и для A_6

Во второй главе рассматривается метод построения универсальной формализованной модели описания аудиовизуальных документов. Проведены исследования и разработаны лексические и грамматические средства языка описания аудиовизуальных документов теле- и радиовещания – ("ДТРВ").

На основе развитого фиксированного способа свертывания текстовой информации в диссертации предложен следующий алгоритм семантического преобразования первоисточника для аудиовизуальных документов:

$$F_i \Rightarrow \mathcal{J}_{1(A)} \vec{\vee} \mathcal{P}_{sem1} \vec{\vee} \mathcal{P}_{sem2}(G) \vec{\vee} \mathcal{P}_{sem3}(\mathcal{L}) \vec{\vee} \mathcal{P}_{sem4}(Q) \vec{\vee} \mathcal{P}_{sem5} \vec{\vee} \mathcal{J}_{2(T)} \ldots /5/$$

где

$\mathcal{P}_{sem1}$ – восприятие первичного документа,

$\mathcal{P}_{sem2}(G)$ – формирование мысленной модели плана задания документа,

$\mathcal{P}_{sem3}(\mathcal{L})$ – выбор лексики языка описания документа,

$\mathcal{P}_{sem4}(Q)$ – выбор грамматических средств языка описания,

$\mathcal{P}_{sem5}$ – построение плана отображения аудиовизуального документа.

Таким образом, семантическое преобразование первичного доку-

мента требует предварительного построения формализованной модели описания, определения лексики и разработки соответствующей грамматики языка описания.

Методика построения формализованной модели сводится к выделению базовых понятий, а затем к их многократному развертыванию в более конкретные смысловые понятия. Глубина развертывания определяется требованием к полноте и точности описания документа. Таким образом, построение плана задания формализованной модели рассматривается как итеративный процесс. Сначала выделяется некоторое множество базовых понятий первого уровня:

$$K_1 = \{K_1^{(1)}, K_2^{(1)}, \ldots K_n^{(1)}\},$$

где $K_i^{(1)}$ – i – тое базовое понятие первого уровня. Далее, $K_i^{(1)} \in K_1$ декомпозируется на базовые понятия второго уровня, затем – третьего уровня и т.д. для каждого последующего уровня. В диссертационной работе достаточно подробно изложен процесс построения универсальной формализованной модели описания аудиовизуальных документов.

Формализованная модель представляет собой неориентированный связанный граф без циклов (дерево) $G = \{X, E\}$, где X – множество базовых понятий, а E – множество грамматических средств, определяющих взаимосвязь лексических единиц базовых понятий различных уровней.

Для описания исследуемых документов теле- и радиовещания предложен новый формализованный язык "ДТРВ". В качестве базового языка взят язык описания документов "Аидос", и на его основе разработан новый язык, отличающийся тем, что в него внесено ряд дополнительных конструкций, которые позволяют более точно описать аудиовизуальный документ в соответствии с морфологическими и семантическими свойствами заданной модели.

Лексика языка "ДТРВ" определяется как суммарная лексика развернутых понятий СЧД и БЧД:

$$\mathcal{L} = \sum_{i=1}^{\ell} \sum_{j=1}^{m} \sum_{t=1}^{n} \mathcal{L}_{СЧД}(K_i^{(t)})_j^* + \sum_{i=1}^{r} \sum_{j=1}^{p} \sum_{t=1}^{q} \mathcal{L}_{БЧД}(K_i^{(t)})_j^* \quad /6/$$

Кроме того, грамматика языка "ДТРВ" расширена по отношению к базовому языку за счет ввода двух дополнительных определителей синтагматических отношений: коэффициент связанности - p и коэффициент перечисления - q.

Коэффициент связанности показывает количество полей - определителей, введенных для характеристики конкретного семантического понятия. Коэффициент перечисления показывает сколько независимых высказываний введено для характеристики базового или развернутого базового понятия.

В диссертации дается оценка разработанного языка "ДТРВ" с позиции существующих классификаций языков и теории формальных грамматик.

При построении подобных систем неотъемлимой частью лингвистического обеспечения является язык запросов (обращения). В этой же главе даны рекомендации построения такого языка и приводится описание грамматики конкретного языка запросов для разработанной системы Гостелерадио СССР.

В т р е т ь е й г л а в е сформулированы и обоснованы основные принципы автоматизированной обработки аудиовизуальной информации.

Исследованы структуры построения основных массивов информационной базы (ОМИБ). Выделены четыре вида структур: 1 - архив, инверсный массив; 2- архив, тезаурус, инверсный массив; 3 - архив, кортежи; 4 - архив, тезаурус, инверсный массив, кортежи.

Так, например, для структуры ОМИБ второго вида, реализованной

во второй системе Гостелерадио СССР, определены зависимости изменения объемов основных массивов от количества документов в фонде системы.

$$V_{тез} = 0{,}4 \cdot 10^2 \Phi$$
$$V_{ин.м} = 1 \cdot 10^2 \Phi$$
$$V_{арх} = 3 \cdot 10^2 \Phi$$

где

Φ - количество документов в фонде,

$V_{тез}$ - объем тезауруса в символах,

$V_{ин.м}$ - объем инверсного массива в символах,

$V_{арх}$ - объем архива в символах.

Для структуры третьего вида исследованы те варианты построения ОМИБ, которые могут быть заданы формализованной моделью, с использованием кортежей, что позволяет учесть семантическую связанность $/ \rho u q /$ между высказываниями БЧД. В общем виде рассмотренная структура ОМИБ может быть представлена следующим образом:

$$ОМИБ = \sum_{i=1}^{n} (T_q)_i + \sum_{j=1}^{m} (T_p)_j + B$$

где

$(T_q)_i$ - кортеж, высказывания в котором находятся в отношении "перечисления" и отнесенные к i - тому базовому понятию,

$(T_p)_j$ - кортеж, высказывания в котором находятся в отношении "связанности" и отнесенные к j - тому базовому понятию,

B - архив.

Четвертый вид ОМИБ дан как рекомендуемая структура построения информационной базы для семантических систем. Данная структура предусматривает одновременное использование основных массивов различных структур:

$$ОМИБ = \sum_{i=1}^{n} (T_q)_i + \sum_{j=1}^{m} (T_p)_j + B + M_{тез} + M_{ин.м}; \quad /8/$$

где

$M_{тез}$ – тезаурус,

$M_{ин.м}$ – инверсный массив,

Рекомендуемая структура раскрывает широкие возможности для выполнения оперативного семантического поиска.

Проведен анализ основных методов поиска, приемлемых для обработки аудиовизуальных документов, и определены условия релевантности по заданным критериям выдачи.

Так:

1. Если критерий выдачи задан, как полное совпадение поискового предписания с планом отображения документа, то условие релевантности будет иметь вид:

$$(\forall(x))[P(x_1, x_2, x_3) \wedge R(x_1, x_2, x_3)]; \quad /9/$$

2. Если критерий выдачи задан, как частное совпадение поискового предписания с планом отображения документа, то условие релевантности будет иметь вид:

$$(\exists(x))[P(x_1, x_2, x_3) \wedge R(x_1, x_2, x_3)]; \quad /10/$$

где $P(x_1, x_2, x_3) \cup R(x_1, x_2, x_3)$ есть 3-х местные предикаты, определенные на одних и тех же множествах: G, L, Q

В результате выполненных исследований были сформулированы

следующие основные положения автоматизированной обработки аудиовизуальных документов:

1. Первичный аудиовизуальный документ / $\mathcal{D}_i$ /, подлежащий автоматизированной обработке, должен принадлежать хотя бы одному из множеств / $\mathcal{D}_\alpha^{**}$, $\mathcal{D}_\beta^{**}$, $\mathcal{D}_\gamma^{**}$ / предложенной классификации. См. /2/.
2. Лингвистическое обеспечение системы определяется алгоритмом семантического преобразования первоисточника. См. /5/.
3. Программное обеспечение системы, кроме своих стандартных функций, предусматривает обработку информации с учетом семантической целостности первичного документа / $\mathcal{D}_i$ / и заданных условий семантической релевантности. См. /9/ и /10/.
4. Структура основных массивов информационной базы системы определяется семантическими отношениями между базовыми или развернутыми базовыми понятиями документа. См. /8/.

В четвертой главе показаны практические результаты функционирования автоматизированной системы обработки аудиовизуальных документов в ГИВЦ Гостелерадио СССР. Весь цикл исследований и проверки полученных результатов был разбит на два этапа.

На первом этапе было выполнено следующее:

- разработана методика выделения базовых понятий в библиографической части аудиовизуальных документов,
- определена лексика языка описания БЧД,
- разработано программное обеспечение системы с учетом "связанности" и "перечисления" высказываний в полях,

- исследована возможность включения разработанной системы в технологический процесс подготовки и выпуска радиопередач,
- проведена численная оценка входных и выходных потоков документов,
- система построена на базе ЭВМ "Минск- 2 3", программы написаны на языке ССК, объем программного обеспечения составил 30 тыс. операторов.

На втором этапе:
- разработана методика выделения базовых понятий в содержательной части аудиовизуальных документов,
- разработана методика формирования плана задания и плана отображения содержательной части аудиовизуальных документов,
- построена универсальная формализованная модель описания аудиовизуальных документов теле- и радиовещания,
- разработан язык описания аудиовизуальных документов телевидения и радиовещания "ДТРВ",
- разработаны программные средства системы, обеспечивающие обработку информации с учетом семантической целостности первичного документа и заданных условий семантической релевантности,
- сформулированы основные положения автоматизированной обработки аудиовизуальных документов с использованием ЭВМ,
- система внедрена на ЭВМ ЕС-1022. Программы напечатаны на языках ПЛ/I и ассемблер. Общий объем программного обеспечения системы составил 100 тыс. операторов, работает в операционной среде ДОС 2.2. Предусмотрена возможность перехода системы на более старшие модели ЕС ЭВМ с использованием операционной среды ОС 6.1.

Правильность предложенных в работе методов подтверждается тем, что в настоящее время обе системы, разработанные на каждом из этапов, успешно функционируют. Первая находится в промышленной эксплуатации с 1974 г., а вторая - в опытной эксплуатации с 1980 г.

В работе приводятся объемно-временные и эксплуатационные характеристики систем и дается их анализ. Сделаны выводы по результатам эксплуатации систем. Выполнен расчет экономической эффективности от внедрения информационной автоматизированной системы обработки аудиовизуальных документов в ГИВЦ Гостелерадио СССР, который составляет около 121 тыс. руб.в год.

ЗАКЛЮЧЕНИЕ

Основные результаты диссертационной работы состоят в следующем:

1. Разработана научная методика автоматизированной обработки аудиовизуальной информации с использованием ЭВМ; сформулированы основные положения этой методики;
2. Разработана классификация аудиовизуальных документов с учетом форм представления семантической информации;
3. Разработан алгоритм семантического преобразования первоисточника;
4. Разработан общий алгоритм семантической обработки аудиовизуальных документов в автоматизированной системе;
5. Построена универсальная формализованная модель и частные рабочие модели описания аудиовизуальных документов;
6. Исследован и разработан формализованный язык описания аудиовизуальных документов "ДТРВ";

7. Исследованы и даны рекомендации построения логической структуры информационной базы;
8. Разработан алгоритм поиска для конкретной информационной базы;
9. Исследован и разработан формализованный язык обращения к системе;
10. Проведены экспериментальные исследования и осуществлено практическое внедрение систем автоматизированной обработки аудиовизуальных документов в ГИВЦ Гостелерадио СССР.

Публикации по теме диссертационной работы:

1. Шатохина С.Н. Некоторые вопросы построения информационно-поискового языка для автоматизированной ИПС, работающей в области радиовещания и телевидения: Тез. докл. междунар. симпозиум ГИ УI ОIРТ Таллин: 1970
2. Кулаков В.Г., Чепурова Л.М., Шатохина С.Н. Технические характеристики и возможности информационно-поисковой системы ТТЦ: Тез. докл. Ш научно-технической конференции ВНИИТРа - М.: 1973.
3. Кулаков В.Г., Шатохина С.Н. Некоторые результаты эксплуатации системы информационного поиска: Тез. докл. IУ научно-техническая конференция ВНИИТРа -М.: 1974.
4. Анисимов Б.В., Белов Б.И., Соломатин Н.М., Кулаков В.Г., Шатохина С.Н. Информационно-поисковая система Телевизионного технического центра: Механизация и автоматизация производства, 1974, № 6, 41-43 с.
5. Шатохина С.Н. Информационно-поисковая система аудиовизуальных документов в Гостелерадио СССР": Тез. докл. семинара "Вопросы разработки и внедрения систем обработки данных на базе ЕС ЭВМ", -М.: МДНТП, 1980.

6. Шатохина С.Н. Разработка информационно-поисковой системы по материалам ТРФ. - Эксплуатационно-технические материалы по телевидению, радиовещанию и кинопроизводству, ГКТР, 1981, № 50 - Разработка и внедрение АСУ Гостелерадио СССР, с 29-34.

7. Соломатин Н.М., Меняев М.Ф., Шатохина С.Н. Принципы автоматизации обработки аудиовизуальных документов. - Труды /МВТУ, 1981, № 365 - Проектирование электронных вычислительных машин и систем - с 98-105

Л-90009 от 2.01.82. Объем 1 п.л. Заказ 42 Тираж 100 экз.
Типография МВТУ им.Н.Э.Баумана.107005 Москва,2-я Бауманская,5