.На правах рукописи

Ибрахим Али Рашид

Разработка методов и программных средств для анализа сходства ациклических структур

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва, 2009

Работа выполнена на кафедре Прикладной математики Московского Энергетического Института (Технического Университета)

Научный руководитель: кандидат технических наук, доцент
Кохов Виктор Алексеевич

Официальные оппоненты: доктор технических наук, профессор
Фальк Вадим Николаевич

кандидат технических наук, доцент
Незнанов Алексей Андреевич

Ведущая организация: Институт вычислительной математики и математической геофизики Сибирского Отделения Российской Академии Наук (ИВМиМГ СО РАН) (Новосибирск)

Защита состоится 09 октября 2009 г. в 16 час. 00 мин.
на заседании диссертационного совета Д 212.157.01 при Московском Энергетическом институте (Техническом Университете)
по адресу: 111250, Москва, Красноказарменная ул., д. 17 (ауд. Г -306)

С диссертацией можно ознакомиться в библиотеке Московского Энергетического Института (Технического Университета).

Отзывы в двух экземплярах, заверенные печатью организации, просьба направлять по адресу: 111250, г. Москва, ул. Красноказарменная, д. 14, Ученый Совет МЭИ (ТУ).

Автореферат разослан « 8 » сентября 2009 г.

Ученый секретарь
диссертационного совета Д 212.157.01
кандидат технических наук,
доцент Фомина М.В.

Общая характеристика работы

Актуальность темы исследований.

Современный этап развития науки характеризуется широким объединением и глубоким взаимопроникновением различных наук, что создает благоприятные условия для постановки и решения сложных научно-технических проблем. К таким проблемам относятся исследования по информационным семантическим системам (ИСС), то есть системам, перерабатывающим осмысленную информацию для достижения целей. В настоящее время исследования, посвященные решению отдельных вопросов функционирования ИСС, активно ведутся как в нашей стране, так и за рубежом. Сравнение и принятие решения – основная семантическая операция.

Важным понятием в ИСС является понятие структуры. Понятие структуры представляется важным с точки зрения классификации существующих и вновь создаваемых форм семантических систем. Принято разнообразие основных структур определять разнообразием графовых моделей систем (ГМС) и их обобщений.

Концепции «подструктура» и «подсистема» являются основой для изучения теорий и теоретических знаний. Концепция подструктуры является такой значимой, например, в химии органических соединений, что для систематизации рассмотрения природы химических структур и подструктур были привлечены методы теории графов и создана химическая теория графов. Анализ сложности структур и разнообразие несходства и подобия в больших объединениях структур сделали необходимым развитие и расширение концепций подструктурной характеризации.

Широкий теоретический и прикладной спектр применения структурного анализа привел к выделению новой дисциплины – прикладной теории графов.

Особую роль методы прикладной теории графов играют именно в развитии информационных технологий (теории трансляции, оптимизации программ, организации сложных структур данных, визуализации данных, построении человеко-машинных интерфейсов и др.). Одним из основных классов задач прикладной теории графов является класс задач различения графов и различения расположения фрагментов графов. История развития методов решения задач различения графов насчитывает более 50 лет. В настоящее время в решении задач распознавания изоморфизма графов, распознавания изоморфного вложения графов и смежных с ними задач достигнут большой теоретический и практический прогресс. Но задачи поиска максимального общего фрагмента двух структур и на основе этого определение сходства структур, изучены намного слабее. Исследования, связанные с учетом расположения фрагментов в структурах актуализировались в конце 90-х годов прошлого века в связи с развитием приложений химической теории графов (*QSAR*-анализа, *QRRR*-анализа и др.), правдоподобных рассуждений, структурного распознавания образов, структурной лингвистики. Эти исследования шли в достаточно узких областях (например, большинство предложенных топологических индексов явно или неявно учитывали

расположение простейших фрагментов) и несистематические. Не существует даже устоявшейся терминологии, пригодной как для теоретических, так и для прикладных исследований. Эта проблема актуализировалась ещё больше после того, как Журавлёв Ю.И. и его ученики с наиболее общих теоретических позиций показали, что при решении задач распознавания выражение глобальных (интегральных) свойств через локальные вполне возможно.

Учёт расположения фрагментов ГМС наполняет новым содержанием отношения «подсистема-надсистема». Задачи различения и сходства расположения фрагментов ГМС обобщают задачи сравнительного анализа ГМС, принимая во внимание надсистему, в которую входят рассматриваемые фрагменты. Решение этих задач позволяет исследовать отношения эквивалентности и толерантности ГМС с учётом расположения и сходства расположения фрагментов, расширяет возможности подструктурной характеризации ГМС. Особенно ярко различия в расположении фрагментов проявляются в симметрии (асимметрии) расположения фрагментов. Учёт симметрии – общеметодологический принцип повышения эффективности компьютерной обработки структурной информации, как в задачах анализа, так и синтеза структур.

Работа продолжает исследования научного руководителя диссертанта – Кохова В.А., которые привели к созданию новой научной дисциплины «структурный спектральный анализ систем» (СС-анализ), и позволяют на единой методологической основе строить достаточно эффективные алгоритмы решения базовых классов задач структурной информатики. Выделены семь основных классов задач СС-анализа, среди которых – класс задач определения сходства и определения сходства расположения фрагментов в ГМС.

Ввиду широкой теоретической и прикладной значимости ациклических структур (деревьев и орграфов без контуров) основное внимание при разработке алгоритмов и программ в диссертации уделено ациклическим структурам.

Целью диссертационной работы является создание методов и программных средств для эффективного решения задач определения сходства *ациклических структур систем* (АС), которые представлены графами-деревьями или ориентированными графами без контуров. Это позволит повысить эффективность компьютерных методов анализа сходства структур систем и их применения при создании новых поколений информационно поисковых систем структурной информации и систем искусственного интеллекта (СИИ) с правдоподобными рассуждениями. Широко использовать их в научных и прикладных исследованиях, связанных со структурным анализом систем.

Для достижения указанной цели в работе решаются следующие задачи:

- Разработка методов построения и исследования инвариантов, характеризующих расположение фрагментов в ациклических структурах с использованием расширяемых базисов структурных дескрипторов.

- Классификация задач определения максимальных общих фрагментов в ациклических структурах и разработка методов их решения.
- Разработка методов решения задач анализа сходства ациклических структур с учетом расположения их фрагментов.
- Разработка методов решения задач анализа сходства расположения фрагментов в ациклических структурах.
- Построение программной подсистемы «Сходство ациклических структур» и её использование в учебном процессе и научных исследованиях.

Научная новизна исследования состоит в следующем:

1) предложены модели, характеризующие ациклические структуры систем с учетом расположения фрагментов, позволяющие:
   a. Расширить и обобщить подструктурный подход к анализу сходства ациклических структур;
   b. Отобразить (визуализировать) любые фрагменты АС в виде цветных вершин модели, что с теоретической точки зрения упрощает анализ *t*-групп для АС, и позволяет внедрять новые информационные технологии в проведение исследований по анализу сходства структур систем;
2) предложены методы и алгоритмы для решения задач анализа сходства ациклических структур, которые могут точно (до орбит *t*-групп) учитывать расположение фрагментов (цепей и путей);
3) предложены методы и алгоритмы для решения задач анализа сходства ациклических структур, которые могут приближенно (до классов эквивалентного расположения фрагментов, полученных на основе инвариантов), учитывать расположение фрагментов (цепей и путей);
4) предложены ЭВМ-ориентированные методы формирования и исследования новых видов отношений эквивалентности и толерантности ациклических структур систем.

В работе показана перспективность и эффективность учёта расположения фрагментов при решении задач определения сходства ациклических структур с использованием вычислительной техники.

Практическая значимость работы заключается в создании методов и на их основе базового программного обеспечения, позволяющего повысить качество и эффективность решения задач структурного анализа систем, связанных с определением их сходства. Результаты важны для приложений структурной информатики и могут быть применены в системах искусственного интеллекта и поддержки принятия решений, в структурном распознавании образов и интеллектуальном анализе данных, семантическом *web*-поиске документов в базах знаний, Интернете и др.

Результаты работы внедрены в учебный процесс МЭИ(ТУ), Государственного университета – Высшей школы экономики и научно-исследовательскую работу кафедры Прикладной математики АВТИ МЭИ (ТУ).

Методы исследований и достоверность результатов. Задачи, поставленные в работе, решаются с помощью методов теории графов, прикладной теории графов, теории групп, теории анализа вычислительной сложности алгоритмов, анализа и построения эффективных алгоритмов и др. В работе существенно использованы результаты, которые получили Кохов В.А, Фараджев И.А., Грызунов А.Б., Ткаченко С.В., Незнанов А.А, Скоробогатов В.А., *J.R. Ullmann, B.D.McKay, G.F. Royle, P.Willett, E.Luks, L.E.Druffel* и др.

Достоверность научных результатов подтверждена теоретическими выкладками, результатами тестирования, а также сравнением полученных результатов с результатами, приведенными в научной литературе.

Реализация результатов. Разработанные программные средства используются в научных исследованиях ИВМиМГ СО РАН, учебном процессе и научно-исследовательской работе кафедры Прикладной математики МЭИ (ТУ), кафедры «Анализ данных и искусственный интеллект» Государственного университета – Высшей школы экономики.

Апробация работы. Основные положения и результаты диссертации докладывались и обсуждались на четырнадцатой международной конференции «Информационные средства и технологии», г. Москва, (2006 г.) и тринадцатой международных научно-технических конференциях студентов и аспирантов «РАДИОЭЛЕКТРОНИКА, ЭЛЕКТРОТЕХНИКА И ЭНЕРГЕТИКА» (г. Москва, 2007г.).

Личный вклад диссертанта. Работа развивает методы структурного спектрального анализа систем для повышения качества и эффективности обработки структурной информации на ПЭВМ.

Диссертантом выполнены.

1) Классификация задач определения максимального общего фрагмента ациклических структур, лежащих в основе системного анализа и развития возможностей методов анализа сходства, использующих максимальные общие фрагменты структур.
2) Разработка базовых моделей, для решения задач определения сходства расположения фрагментов (цепей и путей) в ациклических структурах.
3) Разработка методов решения задач анализа сходства ациклических структур, которые учитывают точное (до орбит t-групп) и приближенное расположение фрагментов (цепей и путей).
4) Исследование разработанных алгоритмов и их реализаций, заключающееся в установлении границ применимости, определении теоретических и экспериментальных оценок вычислительной сложности, сравнении с ранее существующими алгоритмами.
5) Реализация разработанных алгоритмов в виде подсистемы «Сходство ациклических структур» для АСНИ «*GraphModel Workshop*» (*GMW*) и программных средств учебного назначения.

Публикации. Основные результаты диссертационной работы, опубликованы в трех печатных работах, включая одну статью в журнале, рекомендуемом ВАК для публикации основных результатов диссертационных работ.

Структура и объём работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы (105 наименований) и трех приложений. Диссертация содержит 155 страниц машинописного текста.

Содержание работы

Во **введении** обоснована актуальность темы диссертационной работы, поставлены цели и задачи исследований, сформулированы научная новизна и практическая значимость, приведено краткое содержание работы по главам.

В **первой главе** рассматриваются основные подходы к анализу сходства графовых моделей систем. Отмечено, что сдерживающим фактором на пути создания новых поколений информационно-поисковых систем структурной информации, семантическом *web*-поиске документов в интернете, в СИИ с правдоподобными рассуждениями, является отсутствие развитых теорий сходства и инструментальных программных средств для определения и исследования сходства структурированных нечисловых объектов (графов, мультиграфов, гиперграфов). Выделено, что, несмотря на большие успехи, достигнутые в области теории графов, прикладной теории графов и химической теории графов, вопрос разработки ЭВМ-ориентированной методологии для определения и исследования отношений эквивалентности, толерантности и упорядочения графовых моделей систем с учетом расположения фрагментов остаётся открытым и является центральным вопросом структурного анализа систем.

Приведен обзор по основным теоретическим результатам анализа отношений толерантности структур систем. На основе вида *математических пространств* (скалярное, векторное, матричное, теоретико-графовое) приведена систематизация научных работ по анализу сходства структур систем. Выделены 4 подхода к анализу сходства: (1) сходство на основе топологических индексов и индексов сложности графовых моделей систем; (2) сходство на основе использования расширяемых базисов структурных дескрипторов при характеризации структур систем; (3) сходство на основе частично-редуцированного представления графовых моделей структур систем; (4) сходство на основе теоретико-графовых иерархических моделей сложности.

Из анализа подходов сделан вывод, что дальнейшее развитие методов количественного определения сходства становится невозможным без разработки *основ теории структурных инвариантов граф-моделей систем* с акцентом на характеризацию и исследование *структурных инвариантов расположения фрагментов* в графовой модели системы в единой концепции взаимосвязи

«Фрагмент–БазисСтруктурныхДескрипторов–Структура».

Далее приведены основные свойства ациклических структур, выделены примеры областей их применения, актуализирующие анализ сходства орграфов без контуров.

Во **второй главе** приведены теоретические основы подструктурного подхода (ПП) к анализу сходства графов. В ПП выделены две группы методов:

1. Методы, использующие максимальные общие фрагменты пары графов.
2. Методы, использующие базисы структурных дескрипторов.

Основная концепция *первой группы методов* базируется на использовании *максимальных общих фрагментов* двух графов. Этот подход имеет актуальное значение, например, для идентификации *супермолекул* в химической структурной информатике и, следовательно, для определения *структурного сходства между группами графов*, если будут построены эффективные алгоритмы по всему многообразию задач поиска *максимальных общих фрагментов графов*. Во второй группе методов рассматривается поиск наибольших общих фрагментов по двум семействам фрагментов, первое из которых является набором структурных дескрипторов одного графа, а второе – другого графа.

Для определения метрических расстояний между графами используются *подструктурные метрики* (*MCF*-метрики):

1) $D_1(G,H)=|V(G)|+|V(H)|-2(|V(MCS(G,H))|)$;
2) $D_2(G,H)=|V(G)|+|E(G)|+|V(H)|+|E(H)|-2(|V(MCF(G,H))|+|E(MCF(G,H))|)$,

где *MCS* – максимальный общий порожденный подграф, *MCF* – максимальная общая подструктура (частичный подграф максимальный по кардинальному числу).

Для определения сходства графов используется *коэффициент сходства*:

$$MSI(G,H)=\mu(A,B)=|MCF(G,H)|^2/(|G|\times|H|),$$

и *коэффициент несходства* $MDI(G,H)=1.0-MSI(G,H)$.

Результаты вычисления сходства могут быть представлены в виде *матрицы* или *графа попарного сходства* анализируемых АС.

Анализ теоретических основ подструктурного подхода и объемные вычислительные эксперименты по определению сходства (расстояний) графов и АС:

- проанализировано 30399 графов и более 335000 АС;
- решено 5751575 задач для определения максимального общего подграфа для всех пар графов и АС,

привели к выделению следующих недостатков подструктурного подхода.

1. Малая «чувствительность» к различению орграфов, находящихся на расстоянии 1. На расстоянии 1 от заданного орграфа находятся все орграфы, которые можно построить добавлением, либо удалением дуги или одиночной вершины.

2. Жесткая *детерминированность* – полученное значение метрики или коэффициента сходства (несходства) нельзя далее уточнять.

3. Подход не учитывает наличие *значимых фрагментов, их расположение и взаимное расположение в структуре* орграфа.

4. Подход не учитывает качественных характеристик максимальных общих фрагментов (например, их сложность).

5. Узкий диапазон пространства для проведения кластеризации.

С позиций основ теории вычислительной сложности решения задач приведена формализованная постановка *базовых задач* изоморфного пересечения АС в форме задач распознавания свойств.

Задача 1. Наибольший общий слабосвязный подграф двух АС (*MCWS*).

УСЛОВИЕ. Заданы два орграфа без контуров G_1, G_2 и положительное целое число K.

ВОПРОС. Существуют ли $G_1^* \subseteq^S G_1$, $G_2^* \subseteq^S G_2$ такие, что они являются *слабосвязными подграфами*, соответственно в G_1, G_2 и

$$|V(G_1^*)| = |V(G_2^*)| > K \text{ и } G_1^* \approx G_2^*?$$

Задача 2. Наибольший общий слабосвязный фрагмент двух орграфов (*MCWF*).

УСЛОВИЕ. Заданы два орграфа без контуров G_1, G_2 и положительное целое число K.

ВОПРОС. Существуют ли $G_1^* \subseteq {}^f G_1$, $G_2^* \subseteq {}^f G_2$ такие, что они являются *слабосвязными фрагментами*, соответственно в G_1, G_2 и

$$|E(G_1^*)| = |E(G_2^*)| > K \text{ и } G_1^* \approx G_2^*?$$

На рис.1 приведен пример результата решения задачи 1, полученный подсистемой «Сходство ациклических структур» в АСНИ «*GMW*».

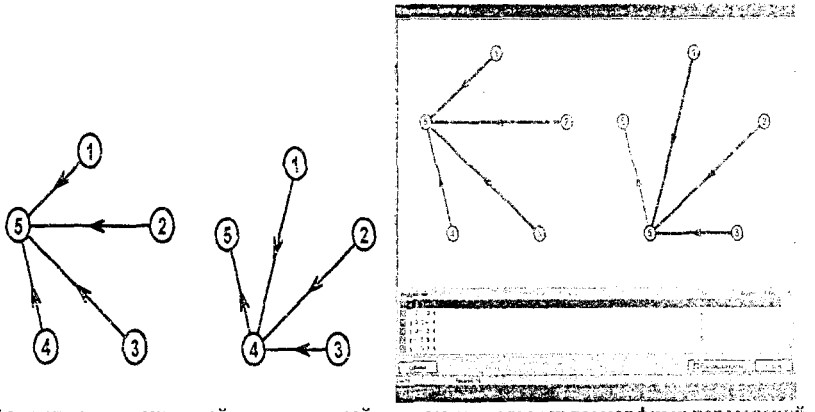

Рис. 1. Вид окна с анимационной демонстрацией всех максимальных изоморфных пересечений по слабосвязным подграфам двух АС в АСНИ «*GMW*»

На основе обобщения результатов анализа задач определения общего фрагмента (*CF*), возникающих в теоретических и прикладных исследованиях, выделен набор из 4 параметров (табл. 1) для классификации задач *CF*.

Таблица 1. Параметры классификации задач определения общего фрагмента орграфов

№	Параметр	Смысл параметра	Виды значений параметра
1	*Part*	Вид соотношения частей в орграфах, которыми являются общие фрагменты	*S-S,S-Fr, Fr-S*
2	*Sl*	Вид результата по количеству общих фрагментов	Одно любое, все, все неизоморфные *CF*
3	*C*	Вид соотношения *CF* по типу связности в АС	Слабо связный – слабо связный; односторонне связный – односторонне связный
4	*Neig*	Окрестность точного решения	Δ=1..3

Часто в прикладных и теоретических исследованиях, например, в химической структурной информатике возникает необходимость поиска решений в заданной окрестности Δ от максимального, то есть

$Neig_1=\max(|V(f)|)-\Delta$; $Neig_2=\max(|E(f)|)-\Delta$, где Δ – целое число.

Таким образом, классификация задач, связных максимальных общих фрагментов для орграфов без контуров, включает 4 параметра:

(1) вид соотношения фрагментов f_1, f_2 как трех частей G_1, G_2;

(2) величину окрестности Δ (Δ=0 – точное решение; Δ>0 – решение в заданной окрестности);

(3) вид результата по числу решений (одно; все);

(4) вид результата по типу связности (k) в анализируемых АС (слабо связные (k=1), односторонне связные (k=2));

В результате выделены 32 вида задач, образующих *первый класс инвариантного ядра задач в анализе ациклических структур*, объединенных в 4 подкласса.

Далее приведен обзор по методам решения задач определения *MCS* и *MCF*, выделены два основных подхода: (1) алгоритмы на основе использования клик в модульном произведении анализируемых графов; (2) алгоритмы на основе метода монотонных расширений частичных решений (ММРЧР), представляющего собой направленный перебор по достройке начального (возможно пустого) решения до максимально возможного с учетом симметрии анализируемых графов, инвариантов, позволяющих отбрасывать бессмысленные достройки, и специфики класса анализируемых ГМС. Приведен сравнительный анализ и выделено, что *наиболее перспективным с точки зрения разработки эффективных алгоритмов является метод монотонных расширений частичных решений* так как:

1. Он позволяет на единой методологической основе разрабатывать алгоритмы решения для всех 32 *базовых задач* с применением разнообразных средств сокращения перебора (использование симметрии и инвариантов расположения фрагментов в анализируемых орграфах).

2. Есть возможность задания частичного решения.

3. Он на единой методологической основе конструктивного перечисления фрагментов в топологии анализируемых орграфов, имеет развитие в методологию решения шести из семи классов базовых задач СС-анализа систем.

4. Алгоритмы, созданные на его основе, работают наиболее эффективно и позволяют задавать ограниченное время для поиска решений. Причем чаще всего точное или близкое к точному решение находится быстро, а затем идет долгий процесс поиска большего по размеру решения.

Алгоритмы определения общих фрагментов, основанные на ММРЧР, и их программные реализации разработаны автором и внедрены в АСНИ «*GMW*». Они позволяют работать с ациклическими структурами, имеющими веса на вершинах, что делает их широкодоступными для применения при решении задач *в различных прикладных областях*, например, при семантическом поиске документов в базах документов.

Приведено описание алгоритма и обоснована его полиномиальность по вычислительной сложности для деревьев при решении задачи определения максимального общего подграфа двух деревьев.

С целью анализа эффективности разработанных алгоритмов *приведены экспериментальные оценки их вычислительной сложности* при определении максимальных общих фрагментов и подграфов для АС. При анализе эффективности алгоритмов *использована оригинальная методика*, включающая разработку и использование генераторов: (1) деревьев с низким, средним и высоким уровнями сложности; (2) орграфов без контуров с низким, средним и высоким уровнями сложности.

В **третьей главе** *предложены граф-модели для характеризации расположения фрагментов в ациклических структурах* с целью разработки новых методов анализа сходства и сходства расположения фрагментов. Приведены основные определения, связанные с анализом t-групп, являющихся индуцированными представлениями группы автоморфизмов вершин ациклических структур, и точно характеризующими расположение фрагментов в этих структурах.

Приведены *оригинальные структурные инварианты*, разработанные автором для ациклических структур. Среди них g-модели, для построения которых требуется нахождения некоторого подмножества помеченных фрагментов АС, которые формируют левую и правую доли (базисы) модели, и последующего анализа попарных пересечений (вложений) помеченных фрагментов. В результате последнего формируются рёбра модели и их веса. При построении g-моделей задавать базисы можно двумя способами:

1. Базис структурных дескрипторов априори задан набором АС и используется процедура поиска всех канонических изоморфных вложений для каждого элемента базиса.

2. Базис структурных дескрипторов строится как подмножество собственных фрагментов АС и используется декомпозиция (разборка) АС на помеченные фрагменты с заданными параметрами. Причём, если мощность базиса СД много меньше числа всех собственных фрагментов АС, имеет смысл сначала найти фрагменты, вкладываемые в АС, с использованием декомпозиции на неизоморфные фрагменты, а затем построить базы помеченных фрагментов с использованием процедура поиска всех

канонических изоморфных вложений. В противном случае лучше использовать декомпозицию с одновременным получением помеченных фрагментов. Показано, что способ с использованием процедуры поиска всех канонических изоморфных вложений *позволяет обрабатывать намного более крупные ациклические структуры.*

В данной работе, автором учтена специфика АС: (1) фрагментами АС являются более простые структуры, чем в обыкновенных и орграфах; (2) разработаны достаточно эффективные алгоритмы распознавания изоморфизма,и изоморфного вложения одной АС в другую; (3) базисы АС состоят из структур без контуров и их надграфов, что существенно сокращает длину базиса структурных дескрипторов.

Показано, что с использованием *g*-моделей можно построить иерархическую систему характеристических инвариантов расположения фрагментов: структурных, матричных, векторных и числовых. Причём не только с целью решения задач различения, но и с целью *развития методов анализа структурной сложности и сходства систем.*

Далее *предложен метод построения цветных* (*cgp*-модели) и *скелетных* (*cgp*-модели) *трансграфов путей* для АС, как класса *g*-моделей, позволяющих:

1) Наиболее эффективно анализировать сходство с учетом расположения путей в АС.

2) Визуализировать расположение цепных фрагментов АС на основе диаграммы АС.

3) Разработать новые методы анализа сходства АС с учетом расположения цепных фрагментов.

Автором для построения *gp*-моделей *построен эффективный алгоритм*, вычислительная сложность которого растёт *практически линейно относительно числа путей* АС. Более того, использование полиномиального по вычислительной сложности алгоритма построения *b*-моделей, на основе *gp*-моделей, позволяющих с разной степенью точности задавать *gp*-модели, и полиномиальный по вычислительной сложности алгоритм поиска их максмального общего подграфа, позволили разработать эффективные методы анализа сходства АС (с числом вершин до 500). Эти методы *впервые учитывают расположение и сходства расположения путей в ациклических структурах.*

Далее подробно рассмотрен класс *g*-моделей как основы для построения систем базисных моделей (*b*-моделей), позволяющих наиболее эффективно решать задачи анализа сходства АС. Приведен переход от класса *g*-моделей к классу *b-моделей АС.* Рассмотрена система подклассов выделенного класса. Каждый из подклассов приводит к *постановкам новых видов задач исследования сходства* АС. На рис. 2 приведен пример трех видов граф-моделей для одного из деревьев с числом вершин 5.

Предложен алгоритм построения *b-моделей* для АС. Основу алгоритма определяет *вычисление числа достроек каждого помеченного фрагмента АС до фрагмента, изоморфного пути (цепи), заданной длины.* Обоснована его

полиномиальная вычислительная сложность для деревьев и приведены экспериментальные оценки вычислительной сложности алгоритмов построения *cgp- и gp-моделей.* Время работы алгоритмов их построения прямо пропорционально количеству добавляемых в граф-модель вершин-путей и поэтому в общем случае *экспоненциально возрастает с увеличением длин путей,* анализируемых АС, симметрия которых отражается в *cgp-моделях.* Результаты анализа оценок алгоритмов показывают высокую эффективность работы процедуры построения трансграфов путей для средних по сложности *планарных орграфов без контуров.* Время работы практически *линейно растёт с ростом числа путей.* Наибольшая эффективность достигается на графах-деревьях и орграфах-деревьях, для которых разработаны специализированные алгоритмы. Для различных семейств ациклических структур приведены оценки эффективности алгоритмов построения *b-моделей.* Показана эффективность построения *b*-моделей для деревьев на основе создания специализированных алгоритмов.

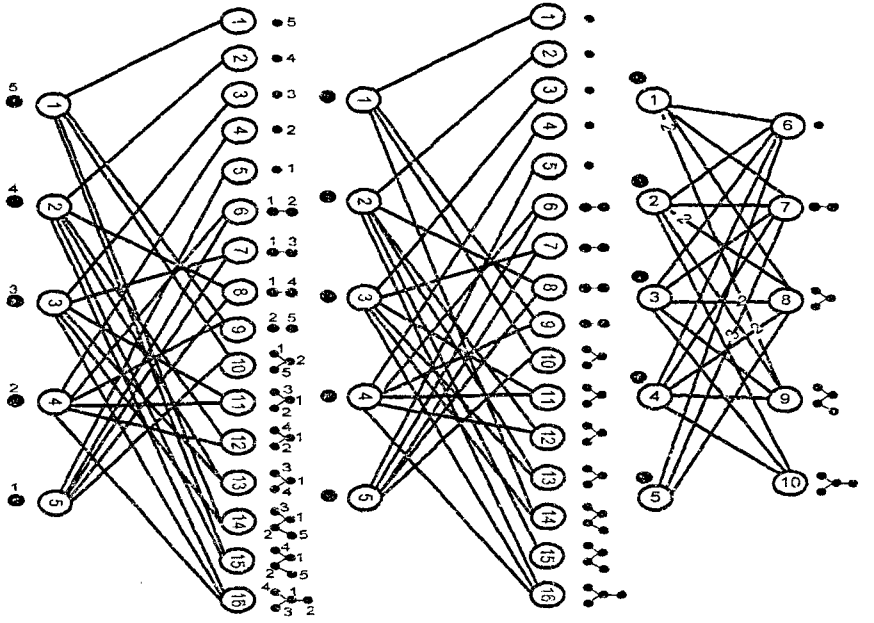

Рис. 2.

Таким образом, на основе предложенных автором *g*-моделей применена наиболее общая концепция подструктурной характеризации для деревьев и ациклических орграфов позволившая:

- разработать расширенный подструктурный подход к анализу сходства деревьев и ациклических орграфов, учитывающий расположение фрагментов в топологии анализируемых графов;

• отобразить (визуализировать) цепные фрагменты дерева или ациклического орграфа в виде цветных вершин *g*-модели, что с теоретической точки зрения упрощает анализ *t*-групп, а с прикладной точки зрения позволяет внедрять новые информационные технологии в проведение исследований и обучение студентов.

В **четвертой главе** рассмотрены 3 метода, расширяющие классический подструктурный подход к анализа сходства и повышающие эффективность его применения для АС:

• расширенный подструктурный подход, впервые учитывающий все *MCS* для АС и качественную информацию (сложность) *MCS* двух АС;

• метод, использующий *MCS* для *g*-моделей (трансграфов путей) АС, что впервые позволяет при анализе сходства учитывать точное расположение фрагментов в ациклических структурах;

• метод, использующий *b*-модели, первые позволяющие *существенно повысить эффективность анализа сходства* АС, учитывать все более и более точно расположение и сходство расположения фрагментов в АС.

В научном аспекте все предложенные методы позволили расширить подструктурный подход к анализу сходства АС. Кроме того, появилась возможность ставить и решать новые классы задач. Эти задачи связаны с анализом расположения и сходства расположения фрагментов в АС, а так же сходства АС с учетом сходства расположения фрагментов заданного типа. Появилась возможность решать новые задачи, связанные с исследованием сходства различных семейств АС по усредненным значениям расстояний между парами АС, которые получены при использовании *g*- и *b*-моделей.

Далее приведена классификация задач анализа сходства двух АС на основе подструктурного подхода. В табл. 2 приведены обозначения базовых задач анализа сходства.

Таблица 2.

N	*N1*	*N2*	Обозначение	Содержание задач
1	1.1	1.1.1	$DG_1{}^{S}\cap_{w}{}^{S}G_2$	Сходство на основе связного *CF* типа слабое-слабое вида П-П
		1.1.2	$DG_1{}^{S}\cap_{wo}{}^{S}G_2$	Сходство на основе связного *CF* типа ОД-ОД вида П-П
		1.1.3	$DG_1{}^{S}\cap_{w}{}^{fr}G_2$	Сходство на основе связного *CF* типа слабое-слабое вида П-Ф
		1.1.4	$DG_1{}^{S}\cap_{wo}{}^{fr}G_2$	Сходство на основе связного *CF* типа ОД-ОД вида П-Ф
		1.1.5	$DG_1{}^{S}\cap_{w}{}^{S}G_2$	Сходство на основе связного *CF* типа слабое-слабое вида Ф-П
		1.1.6	$DG_1{}^{fr}\cap_{wo}{}^{S}G_2$	Сходство на основе связного *CF* типа ОД-ОД вида Ф-П
		1.1.7	$DG_1{}^{fr}\cap_{w}{}^{fr}G_2$	Сходство на основе связного *CF* типа слабое-слабое вида Ф-Ф
		1.1.8	$DG_1{}^{fr}\cap_{wo}{}^{fr}G_2$	Сходство на основе связного *CF* типа ОД-ОД вида Ф-Ф
	1.2	1.2.1	$DG_1{}^{S}\cap_{\Delta w}{}^{S}G_2$	Сходство на основе связного Δ-*CF* типа слабое-слабое вида П-П
		1.2.2	$DG_1{}^{S}\cap_{\Delta wo}{}^{S}G_2$	Сходство на основе связного Δ-*CF* типа ОД-ОД вида П-П
		1.2.3	$DG_1{}^{S}\cap_{\Delta w}{}^{fr}G_2$	Сходство на основе связного Δ-*CF* типа слабое-слабое вида П-Ф
		1.2.4	$DG_1{}^{S}\cap_{\Delta wo}{}^{fr}G_2$	Сходство на основе связного Δ-*CF* типа ОД-ОД вида П-Ф
		1.2.5	$DG_1{}^{S}\cap_{\Delta w}{}^{S}G_2$	Сходство на основе связного Δ-*CF* типа слабое-слабое вида Ф-П
		1.2.6	$DG_1{}^{fr}\cap_{\Delta wo}{}^{S}G_2$	Сходство на основе связного Δ-*CF* типа ОД-ОД вида Ф-П
		1.2.7	$DG_1{}^{fr}\cap_{\Delta w}{}^{fr}G_2$	Сходство на основе связного Δ-*CF* типа слабое-слабое вида Ф-Ф
		1.2.8	$DG_1{}^{fr}\cap_{\Delta wo}{}^{fr}G_2$	Сходство на основе связного Δ-*CF* типа ОД-ОД вида Ф-Ф

Подробно рассмотрен первый из новых методов анализа сходства. Он расширяет возможности подструктурного подхода на основе использования дополнительной к размеру *MCS* информации (число всех максимальных изоморфных пересечений двух АС, число канонических максимальных изоморфных пересечений). Далее используются характеристики, позволяющие учитывать *качественный состав MCS* (например, индексы и вектор-индексы сложности *MCS*). На основе объемных вычислительных экспериментов по данному методу сделан вывод, что следует разрабатывать и другие методы к анализу сходства АС, желательно с большим значением чувствительности различения пар АС, анализируемых на сходство, чем при первом методе.

Далее предложена и рассмотрена система методов анализа сходства АС с использованием трансграфов путей, построенных с учетом длин путей (0-1; 0-2; 0-3; ...;(0-(p–1))) исходной АС. Эти методы, используя поиск максимального общего подграфа для трансграфов, впервые позволяют учитывать расположение путей разных длин в АС. Появилась возможность анализировать: (1) тенденции изменения сходства и выделять границы стационарности значений сходства при увеличении длин путей; (2) сходство АС относительно выделенного фрагмента в АС. Эти возможности программно реализованы автором в виде подсистемы «СС-анализ ациклических структур» в АСНИ «*GMW*».

Данный подход на основе системы g-моделей позволил ввести и использовать *новые характеристики для исследования сходства* АС: (1) среднее расстояние между АС относительно страт g-моделей; (2) индекс среднего сходства между АС относительно страт g-моделей.

С использованием новых характеристик появляется возможность строить и анализировать следующие графы: (1) граф сходства АС со средними расстояниями относительно страт g-моделей; (2) граф сходства АС со средними значениями индексов сходства между графами относительно страт g-моделей.

Таким образом, использование системы страт g-моделей приводит к новым аспектам *более универсального и более точного анализа сходства* на основе учета расположения фрагментов, которые интересуют исследователя.

Третий из новых предложенных методов использует построение b-моделей и поиск их максимальных общих фрагментов. b-модели представлены матрицами достроек помеченных фрагментов АС до фрагментов, изоморфных элементам базиса путей, поддеревьев и др. Они позволяют при расширении базиса все более и более точно характеризовать расположение фрагментов в АС. На основе их максимального общего фрагмента легко ввести меру попарного расстояния или сходства АС. Приведен алгоритм построения b-моделей, позволяющих характеризовать расположение путей в базисе путей. На основе объемных вычислительных экспериментов показано, что исследование сходства на основе b-моделей *является наиболее эффективным и чувствительным методом из всех рассмотренных методов*, основанных на подструктурном подходе. Данный метод позволяет эффективно анализировать сходство АС с числом вершин до 400-600 и при этом учитывать расположение

путей в АС. Для деревьев и корневых растущих ордеревьев число вершин возрастает до 5000.

На основе объемных вычислительных экспериментов (АС с числом вершин от 3 до 8 (более 240000)) приведены результаты сравнительного анализа по четырем методам анализа сходства. Рассмотрены: (1) задачи определения сходства расположения фрагментов в АС и методы их решения на основе *b*-моделей; (2) метод решения задачи определения сходства расположения фрагментов на основе *b(g)*-моделей, позволяющий при некоторой потере точности существенно повысить эффективность анализа сходства.

Показана эффективность решения задач анализа сходства по классическому подструктурному подходу и расширенному подходу на средних по сложности деревьях и корневых растущих ордеревьях. Приведены результаты определения экспериментальных оценок вычислительной сложности алгоритмов решения задач анализа сходства АС по оригинальной методике (для различных семейств АС с низким, средним и высоким уровнями сложности в базисе путей), показывающие границы их применимости. На рис.3 приведены экспериментальные оценки при анализе сходства в классе графов-деревьев по расширенному подструктурному подходу (по оси X – число вершин, анализируемых деревьев, по оси Y – время в миллисекундах).

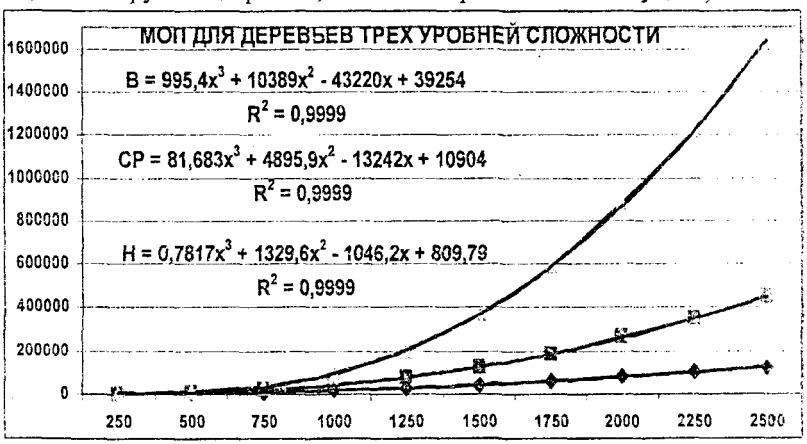

Рис. 3.

В **приложении 1** приведены основные определения и описания разработанных алгоритмов с анализом их вычислительной сложности.

В **приложении 2** рассмотрена подсистема структурного поиска информации в АСНИ «*GMW*», использующая разработанные алгоритмы и программы. Особенностями реализации структурного поиска в подсистеме *GMW* являются: (1) поддержка всех видов точной и приближенной структурной характеризации текста документа; (2) поиск по ранее вычисленному и сохранённому фильтру базы структур; (3) возможность проведения многоэтапного (уточняющего) поиска; (4) использование *b*-моделей при

реализации расширенного подхода к анализу сходства АС; (5) открытый программный интерфейс для улучшения существующих и добавления новых поисковых механизмов.

Способ задания критерия поиска для каждого этапа поиска разбит на три части (что обеспечивает гибкость и эффективность реализации, а также упрощает программные интерфейсы для расширения средств поиска):

1. Используемый инвариант АС, то есть вычисляющее его программное расширение в АСНИ и параметры его запуска. При этом, если для базы АС уже вычислялся данный инвариант, то его повторное вычисление не выполняется, а только сравнивается его значение со значением для орграфа-шаблона. В качестве инварианта может выступать сама АС (классический подструктурный поиск), *g*-модель или *b*-модель АС, метаинформация об АС (название, описание и другие атрибуты).
2. Метод сравнения инвариантов и получения коэффициента сходства АС, также в виде программного расширения и параметров его запуска.
3. Ограничения на значения инварианта и коэффициента сходства.

Исходная база АС *не изменяется* в процессе иерархического поиска. Механизм поиска изменяет только связанную с базой АС *базу данных результатов экспериментов* (БДР). Основными объектами БДР при этом являются фильтры базы АС (определяющие области поиска), таблицы со значениями инвариантов (связанные с ГМС отношением 1:1) и таблицы попарного сходства. Поиск заканчивается либо созданием фильтра базы АС, либо представлением результатов для интерактивного уточнения (только на последнем этапе).

Проведение многоэтапного иерархического поиска включает: (1) выбор типа инвариантов, используемых в процессе поиска; (2) расчёт значений инвариантов на данном этапе; (3) задание критериев сравнения (выбор ПР) и выполнение поиска с созданием нового фильтра; (4) повторение выполнения пунктов 1-3 для новой области поиска (фильтра) или интерактивный анализ результатов на последнем этапе.

Многокритериальный структурный поиск реализуется в АСНИ путём выбора на очередном этапе другой АС в качестве шаблона поиска.

В качестве шаблона поиска может выступать как специально созданная (в редакторе шаблонов) АС, так и одна из АС исследуемой базы.

Эффективная реализация базовых алгоритмов сравнения АС позволяет за приемлемое время (менее минуты на одну АС на компьютере с процессором *AMD Athlon* 64 3000+ и 1 ГБ оперативной памяти) обрабатывать АС следующего порядка: (1) при полной идентификации – до 1500 вершин; (2) при частичной идентификации – до 700 вершин; (3) при подструктурном сходстве – до 80 вершин; (4) при приближённой характеризации время определяется видом используемых *b*-моделей. Например, при использовании полного структурного спектра в базисе путей (цепей) малых длин (от 0 до 6) – до 500 вершин. Так например, при поиске в базе из 237318 АС при использовании числа цепей с длинами от 0 до 3 было получено 11526 АС при границе сходства

95%. При использовании системы вида

$P_{0\text{-}2}(95\%) \rightarrow P_{0\text{-}2}(98\%) \rightarrow P_{0\text{-}4}(98\%) \rightarrow P'_0 \subseteq P_{0\text{-}2}(95\%) \rightarrow P'_0 \subseteq P_{0\text{-}4}(92\%) \rightarrow P'_{0\text{-}3} \subseteq P_{0\text{-}4}(98\%)$

11526 → 2143 → 183 → 43 → 19 → 3

в конечном итоге было выделено 3 АС при границе сходства 98%.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИОННОЙ РАБОТЫ

В диссертации разработаны методы, алгоритмы и программы для решения задач определения сходства АС и сходства расположения фрагментов в АС. Предложенные методы ориентированы на решение комплекса задач, связанных с повышением эффективности и с расширением интеллектуальных возможностей современных компьютерных систем поиска и обработки структурной информации.

К основным результатам работы следует отнести:

1. Из анализа научных работ и основных результатов по анализу отношений эквивалентности и толерантности графовых моделей систем показано, что дальнейшее развитие методов количественного определения сходства становится невозможным без использования единой концепции взаимосвязи *«фрагмент–БСД–структура»*. Отличительная особенность концепции заключается в учете *расположения фрагментов*, что приводит к получению *более чувствительной меры сходства и расширении возможностей* классического подструктурного подхода к анализу сходства структур систем.

2. Ввиду широкой научной и прикладной значимости выделен класс ациклических структур как наиболее перспективный для построения эффективных алгоритмов анализа сходства и их применения при семантическом поиске текстовых документов в базах документов и Интернете.

3. С целью расширения классического подструктурного подхода к анализу сходства АС рассмотрены задачи определения максимальных общих фрагментов для АС. Предложена классификация задач, на основе которой выделено 32 вида (4 подкласса) задач, образующих инвариантное ядро задач в структурном анализе систем.

4. На основе анализа работ по двум основным подходам к решению задач определения общих фрагментов графов выделен метод монотонных расширений частичных решений, как наиболее перспективный для разработки эффективных алгоритмов, и дающих возможность получать точный или близкий к точному результат за ограниченное время поиска решения. На его основе разработаны алгоритмы и программы (в среде *Delphi*, расширения выполнены в виде *DDL* библиотеки) для решения базовых задач определения максимального общего фрагмента двух АС и определены экспериментальные оценки вычислительной сложности алгоритмов по оригинальной научной методике.

5. Предложен метод расширения возможностей классического подструктурного подхода на основе использования более полной и, наряду с количественной, (число максимальных изоморфных (канонических

изоморфных) пересечений, также и качественной информации (индексы сложности общих фрагментов).

6. На основе разработки и использования двух классов структурных инвариантов графов (*g*- и *b*-модели) применена обобщенная концепция подструктурной характеризации для ациклических структур, позволившая:

- разработать обобщенный подструктурный подход к анализу сходства АС, учитывающий расположение фрагментов (путей, цепей);
- создать систему методов иерархического уточнения результатов анализа сходства АС на основе расширяемых базисов путей (цепей для деревьев);
- создать достаточно эффективные алгоритмы анализа сходства и повысить размер анализируемых на сходство АС;
- отобразить (визуализировать) любые фрагменты АС в виде цветных вершин *g*-модели, что с теоретической точки зрения упрощает анализ *t*-групп, а с прикладной точки зрения позволяет внедрять новые информационные технологии в проведение исследований и обучение студентов по исследованию сходства структур систем.

7. Разработаны алгоритмы для построения *g*- и *b*-моделей. Приведены экспериментальные оценки вычислительной сложности построения моделей. Установлено, что на основе использования *b*-моделей и *g(b)*-моделей можно достаточно эффективно анализировать сходство АС с существенным расширением пространства кластеризации АС (по сравнению с подструктурным подходом) и числами вершин до 500, в то время как подструктурный подход позволяет анализировать сходство АС с числом вершин до 50.

8. Создана программная подсистема «Сходство ациклических структур» в рамках АСНИ «*GMW*». Она используется в учебном процессе и научных исследованиях студентов МЭИ (ТУ) при изучении базовой дисциплины «Информатика», раздел «Основы структурной информатики»» и спецдисциплин «Теория графов и комбинаторика», «Дискретная математика», «Анализ и проектирование эффективных алгоритмов», студентов факультета бизнес-информатики ГУ-ВШЭ, научных исследованиях кафедры прикладной математики МЭИ (ТУ), кафедры анализа данных и систем искусственного интеллекта Государственного университета – Высшей школы экономики и Института вычислительной математики и математической геофизики СО РАН.

9. Проведены объемные вычислительные эксперименты с использованием АСНИ «*GMW*», которые позволили получить оригинальные научные результаты определения сходства АС с учетом расположения фрагментов.

10. Предложенные в работе методы, алгоритмы и программные средства позволили создать подсистему иерархического поиска структурной информации в больших базах АС. Они ориентированы в первую очередь на применение при разработке новых поколений ИПССИ и СИИ с правдоподобными рассуждениями. Позволяют впервые решать следующие задачи анализа сходства: (1) сходство АС с учетом расположения классов эквивалентно расположенных путей; (2) сходство АС с учетом расположения

орбит *t*-групп, точно характеризующих расположение путей; (3) сходство АС с учетом сходства расположения классов эквивалентно расположенных путей в АС; (4) сходство расположения путей между классами эквивалентного расположения путей.

Список работ, опубликованных по теме диссертации:

1. **Кохов В.А., Ибрахим А.Р., Кохов В.В.** Система моделей для анализа сходства графов с учетом расположения цепей // Вестник МЭИ, №5, 2009. – С. 5-10.

2. **Кохов В.А., Горшков С.А., Ибрахим А.Р, Джасим М.Р.** АСНИ «Мастерская граф-моделей»: подсистема структурного спектрального анализа деревьев. Труды международной конференции «Информационные средства и технологии», МФИ-2006, Т2. Москва, МЭИ, 2006. – С. 104-108.

3. **Ибрахим А.Р, Джасим М.Р., Кохов В.А.** Программный комплекс для структурного спектрального анализа ациклических графов. Труды тринадцатой международной научно-технической конференции «Радиоэлектроника, электротехника и энергетика», Т1. М., МЭИ, 2007. – С. 360-361.

Подписано в печать 1.09.09г. Зак. 190 Тир. 100 П.л. 1,25

Полиграфический центр МЭИ (ТУ)
111250, г. Москва, ул. Красноказарменная, д. 13