

УДК 519.1: 004.42

ЗАДАЧИ АНАЛИЗА СТРУКТУРНОГО СХОДСТВА СИСТЕМ И МЕТОДЫ ИХ РЕШЕНИЯ

В.А. Кохов (viktorkokhov@rambler.ru)

А.А. Незнанов (aneznanov@hse.ru)

С.В. Ткаченко (svt@richview.com)

Государственный университет –
Высшая школа экономики, Москва

Рассматривается современное состояние дел в области структурного анализа систем. Обсуждаются общие подходы к определению структурного сходства и методы анализа структурного сходства графовых моделей систем. Демонстрируется реализация обсуждаемых методов в различных подсистемах АСНИ «*Graph Model Workshop*».

Введение

Исследование структурного сходства непрерывно повышает свою актуальность, что определяется потребностями структурного анализа систем в различных областях (интеллектуальный анализ данных, химическая структурная информатика, структурная лингвистика, структурная социология и другие). В докладе рассматривается классификация взглядов, подходов и методов анализа структурного сходства, а также реализация методов в виде программных средств.

Уточним, что структура системы моделируется *графовой моделью системы* (ГМС). ГМС может принадлежать любому классу графовых моделей (от обыкновенных графов до взвешенных гиперграфов), что определяется предметной областью. Отметим резкое возрастание числа объектов исследования, моделируемых ГМС, особенно в социально-экономических науках [Jackson, 2008]. Развитие методов искусственного интеллекта актуализировало развитие методов структурного анализа в связи с расширением спектра структурных форм представления знаний [Chein et al, 2009] и развитием таких прикладных областей как структурное распознавание образов. Например, всё большее применение находят понятийные графы и решётки формальных понятий [Hitzler et al, 2009]). При анализе сходства становится важным не только сравнение классических топологических характеристик структур, но и комплексный

учёт класса структур, весов вершин и рёбер/дуг, ограничений на взаиморасположение фрагментов и др. От классических мер структурного сходства переходят к рассмотрению более общих отношений толерантности и подобия. Для удобства исследований вводят различные промежуточные представления исходных ГМС, сохраняющих и рельефно акцентирующих необходимые характеристики.

Наиболее общая постановка задачи для двух структур звучит так: пусть даны две ГМС G_1 и G_2 , требуется количественно оценить степень сходства между ними. Обычно сходство выражают либо мерой сходства (чем больше мера, тем более похожи ГМС друг на друга, стандартный диапазон значений $(0, 1]$), либо коэквивалентной мерой различия (чем больше мера, тем менее похожи ГМС, стандартный диапазон значений $[0, 1)$). Меры сходства/различия легко строить путём нормировки абсолютных значений расстояний. Естественнo потребовать, чтобы хорошая мера различия обладала свойствами метрики. Определение меры сходства позволяет перейти к наиболее практически значимым задачам кластеризации множества ГМС, нечёткого поиска в базе ГМС и др.

1. Подходы к определению структурного сходства систем

Выделим два основных взгляда на процедуры определения структурного сходства.

Подструктурный взгляд (рис. 1) акцентирует внимание на проблеме поиска максимальных общих фрагментов (МОФ) ГМС, которая является *NP*-полной теоретико-графовой проблемой. Из большой группы методов можно выделить *классический подструктурный* подход и *расширенный подструктурный* подход.

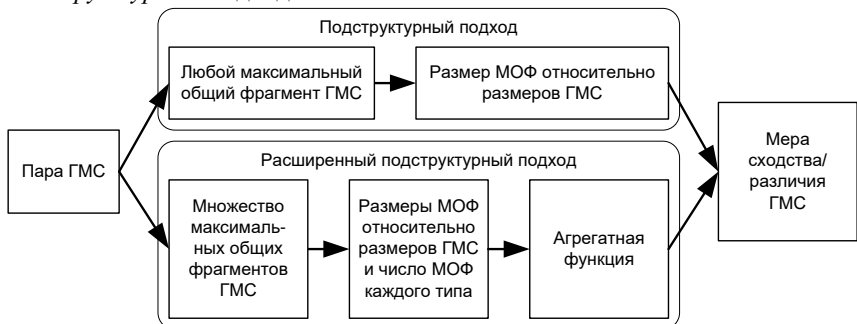


Рис. 1. Подструктурный взгляд на определение структурного сходства пары ГМС

Стандартное определение меры сходства по подструктурному подходу для МОФ в смысле подграфа:

$$MSI1(G_1, G_2) = |V(MCS(G_1, G_2))|^2 / |V(G_1)| |V(G_2)|,$$

а для МОФ в смысле частичного подграфа:

$$MS1(G_1, G_2) = |E(MCS(G_1, G_2))|^2 / |E(G_1)| |E(G_2)|.$$

Эти меры интерпретируются очевидным образом, но сложно вычисляются и обладают низкой дистинктивной способностью. Расширение подструктурного подхода заметно повышает дистинктивную способность, но увеличивает как временную, так и емкостную вычислительную сложность.

Структурно-характеристический взгляд (рис. 2) пытается независимо рассматривать ГМС, чтобы получить промежуточные представления, выявляющие некоторые свойства ГМС, для которых уже существуют адекватные процедуры сравнения. Сюда можно отнести и популярный в последнее время подход на основе *графовых ядер (graph kernels)* [Gartner, 2009].



Рис. 2. Структурно-характеристический взгляд на определение структурного сходства пары ГМС

Структурно-характеристический подход может предполагать учёт следующих характеристик ГМС.

1. Метрические (например, максимальной длины цепей/путей);
2. Спектральные (например, факта вхождения в ГМС фрагментов определённых типов и числа таких вхождений);
3. Характеристики симметрии (например, числа орбит вершин или более сложных фрагментов);
4. Характеристики взаимного расположения фрагментов (например, наличия непустых пересечений помеченных фрагментов определённых типов).

Многообразие структурных характеристик ГМС, которые могут быть представлены в простом для сравнения виде, чрезвычайно велико. В общем виде они изучаются теорией структурной сложности. Модели структурной сложности, которые выявляют те или иные характеристики, используются давно и плодотворно. Достаточно указать на применение в

хим- и биоинформатике разнообразных топологических индексов [Devillers et al, 2000]. Несмотря на значительные успехи, мы находимся, по сути, в начале пути осознания междисциплинарной связи структурной сложности и сходства систем.

Основными формами представления моделей структурной сложности являются число, вектор и матрица. Однако интересно проанализировать, что получится в результате доведения структурно-характеристического подхода до предела выразительности с получением графовых моделей в качестве промежуточного представления. Очевидно, что к полученным структурным моделям структурной сложности (СМСС) может быть применён подструктурный подход. Такая двухэтапная схема получила название обобщённого подструктурного подхода (ОПП) (рис. 3).



Рис. 3. Обобщённый подструктурный подход к анализу структурного сходства пары ГМС

Предлагаемая реализация ОПП базируется на предложенной Коховым В.А. [Кохов, 2002] и развиваемой авторами работы системе структурных инвариантов, названных *g*-моделями ГМС. Модели универсальны в смысле произвольности исследуемых ГМС и произвольности структурных дескрипторов. В системе выделены подклассы моделей, обладающих свойствами, позволяющими вычислительно эффективно решать проблему поиска их МОФ. Это позволяет говорить о практической применимости обобщённого подструктурного подхода. К достоинствам ОПП на основе *g*-моделей можно отнести комплексность характеристики ГМС с учётом взаимного расположения произвольных фрагментов и многообразие подмоделей с различными свойствами. Недостатком является сложность настройки на предметную область, что, однако является общим недостатком всех универсальных методов анализа структурного сходства.

2. Анализ сходства на практике

При решении прикладных задач число ГМС, которое необходимо исследовать, зачастую достаточно велико и может достигать десятков миллионов. В этом случае задача полного анализа попарного сходства всего множества графов лишается смысла. Как следствие, в некоторых случаях формулируется задача нечёткого поиска как задача отбора ГМС, наиболее сходных с заданной. Если же необходимо провести кластеризацию или построить и визуализировать граф сходства

применяются локально-оптимальные («жадные») алгоритмы. Это делает невозможным применение подструктурного подхода и ещё больше актуализирует проблему выбора адекватных характеристик при использовании структурно-характеристического подхода. Это объясняет также популярность подхода на основе графовых ядер.

Понимание особенностей описания структурных характеристик позволило создать методологию применения ОПП для всех классов задач анализа сходства: 1) пары ГМС; 2) расположения пары помеченных фрагментов в ГМС; 3) пары ГМС с учётом расположения помеченных фрагментов; 4) пары ГМС с учётом *сходства* расположения помеченных фрагментов. Сходство расположения помеченных фрагментов – исключительно значимая характеристика структуры системы. Эта характеристика стала приниматься во внимание исследователями не так давно (в конце 80-х годов) и сравнительно редко ввиду неразработанности математических моделей учёта расположения помеченных фрагментов, интересующих исследователя. Особую трудность представляет построение эффективно вычислимых моделей. В области хим- и биоинформатики изменение взаимного расположения фармакофоров (при неизменности всех остальных характеристик) приводит к резкому изменению активности. Перемещение подгруппы в социальной сети также резко меняет её свойства.

Предлагаемая методология разделяет процессы исследования предметной области (требующегося для уточнения форматов ГМС и настройки параметров g-моделей [Незнанов и др., 2008]) и собственно анализа сходства. Параметризация заключается в выборе:

- 1) структурных дескрипторов;
- 2) априорных оценок их сложности;
- 3) типа подмодели структурной сложности по критериям выразительности и эффективности вычисления;
- 4) типа процедуры сравнения моделей или меры сходства при анализе СМСС;
- 5) набора учитываемых весов вершин и рёбер ГМС;
- 6) вида следующего этапа исследований (повторение анализа сходства по ОПП на суженном множестве ГМС, кластеризация выбранным методом; упорядочение; статистический анализ характеристик моделей и т.д.).

Большое число параметров ОПП актуализировало необходимость создания инструментов оценки влияния вносимых изменений на поведение получаемого отношения сходства. Это было проделано С.В. Ткаченко путём реализации подсистемы, позволяющей анализировать изменения в характере сходства при варьировании базиса структурных дескрипторов и выбирать наилучшие подбазисы.

3. Программная реализация методов анализа сходства

Все рассматриваемые подходы к анализу структурного сходства реализованы в АСНИ «*Graph Model Workshop*» (GMW), которая развивается с 1997 года и предназначена для проведения научных исследований и решения прикладных задач в области структурного анализа (рис. 4). Текущая версия GMW 4.0 обладает зрелой архитектурой, построенной на модульном принципе, а её ядро поддерживает три прикладных программных интерфейса, что позволяет безболезненно и с минимальными трудозатратами расширять функциональность АСНИ.

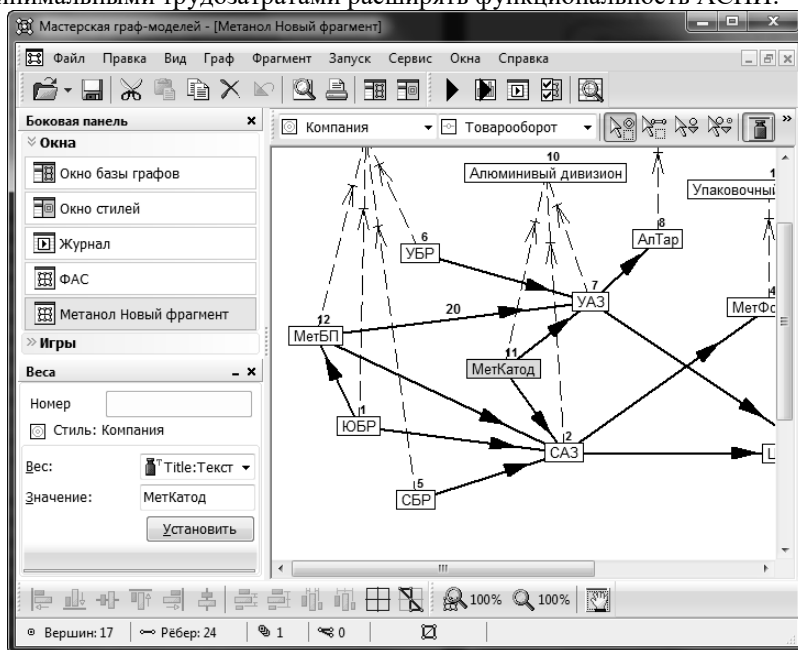


Рис. 4. Главное окно GMW с открытым визуальным реактором ГМС

Универсальная постановка задачи анализа сходства ГМС позволила к настоящему времени создать набор подсистем АСНИ «*Graph Model Workshop*», предназначенных для:

- 1) многоэтапного анализа сходства ГМС на основе произвольных мер сходства с возможностью визуализации и обработки результатов;
- 2) исследования отношений структурной эквивалентности, сходства и подобия;
- 3) реализации обобщённого подструктурного подхода;
- 4) исследования роли симметрии расположения фрагментов в определении сходства их расположения;

5) нечёткого структурного поиска в базах ГМС.

Хотя впервые результаты применения обобщённого подструктурного подхода в *GMW* докладывались в 2004 году [Кохов и др., 2004], к 2009 году этот подход был реализован в полном объёме на всех своих уровнях (уровни ОПП связаны с уровнями выразительности и сложности подмоделей *g*-модели). Анализ сходства можно проводить как в исследовательском режиме, когда промежуточные структурные модели сложности становятся обычными базами ГМС (рис. 5), так и в прикладном режиме, когда все промежуточные модели хранятся в базе данных результатов экспериментов АСНИ (БДР).

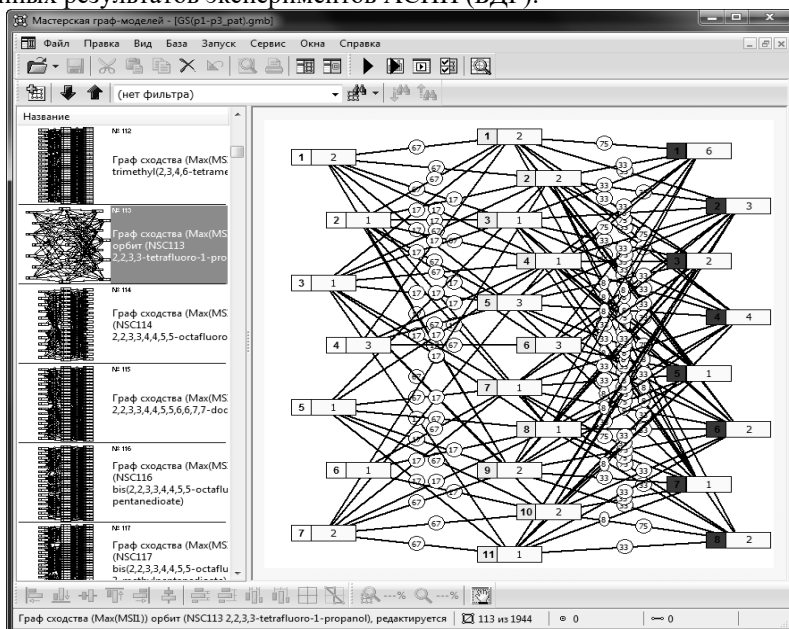


Рис. 5. База GSO-моделей – промежуточного представления характеристик симметрии расположения фрагментов при анализе сходства по ОПП

На основе ОПП были проведены объёмные вычислительные эксперименты в области химической структурной информатики. Были проанализированы базы Национального института исследования рака (более 270 000 ГМС) и базы Российского онкологического научного центра (более 700 ГМС). Затраченное машинное время – более 400 часов. На рис. 6 показан пример визуализации графа сходства на последнем этапе экспертного анализа. Всего проанализировано 12 вариантов параметризации иерархических *GSO*-моделей в базисах цепей и циклов [Незнанов и др., 2008]. Этот класс моделей обладает свойством выявления полной информации о симметрии расположения помеченных фрагментов

при относительно небольшом размере (числе вершин модели). Иерархичность позволяет уменьшить число рёбер модели и повысить эффективность поиска МОФ.

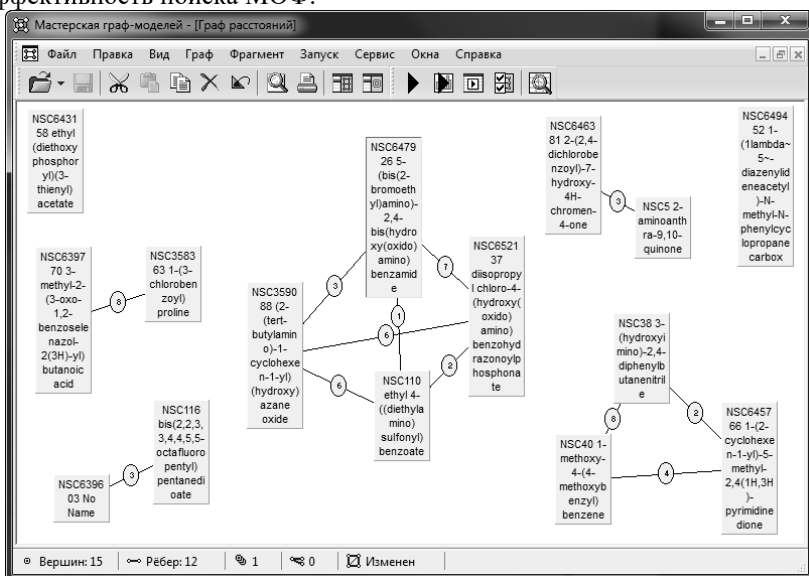


Рис. 6. Граф сходства 15 молекулярных графов базы NCI оставшихся на последнем этапе QSAR-анализа (изображены рёбра, соединяющие графы, сходные более чем на 90%)

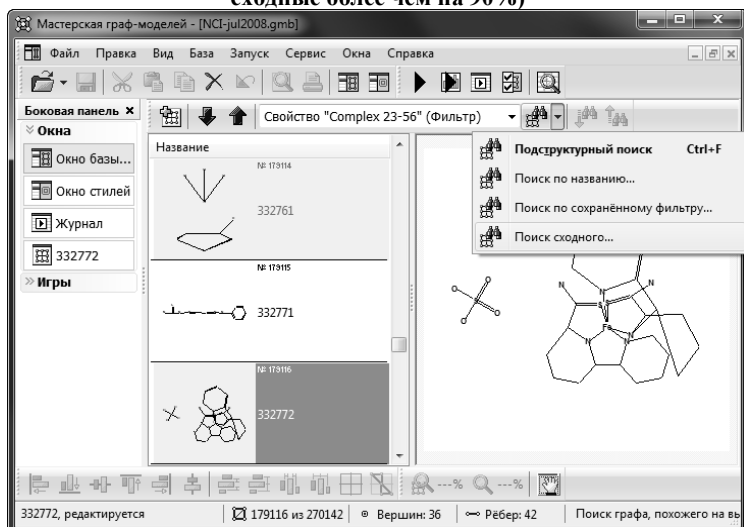


Рис. 7. База ГМС в GMW, открыто меню выбора режима поиска

Результаты экспериментов позволяют утверждать о возможности всестороннего и глубокого исследования сходства семейств молекулярных графов, для которых классический подструктурный подход не позволяет проводить *QSAR*-анализ вовсе, а классические топологические индексы вырождаются.

Структурный поиск в базах ГМС реализован в *GMW* также по модульной схеме. На рис. 7 показано меню поиска в базе ГМС. При выборе пункта «Поиск сходного...» можно задействовать любые расширения АСНИ, которые вычисляют меры сходства. Пространство поиска можно последовательно сужать, так как результатом поиска может быть фильтр базы ГМС. Ядро АСНИ позволяет в процессе поиска использовать информацию, ранее сохранённую в БДР.

Единый подход к решению задач анализа структурного сходства позволил создать оригинальную методологию и программные средства, используемые в областях интеллектуального анализа данных и химической структурной информатики, требующие оригинальных определений сходства и компьютерной поддержки.

Список литературы

- [Chein et al, 2009] Michel Chein, Marie-Laure Mugnier: Graph-based Knowledge Representation, Springer, 2009, XIV.
- [Devillers et al, 2000] Topological Indices and Related Descriptors in QSAR and QSPAR, edited by James Devillers and Alexandru T Balaban, CRC Press, 2000.
- [Gartner, 2009] Thomas Gartner: Kernels For Structured Data, World Scientific Publishing Company, 2009.
- [Hitzler et al, 2009] Conceptual Structures in Practice, edited by Pascal Hitzler and Henrik Scharfe, Chapman & Hall, 2009.
- [Jackson, 2008] Matthew O. Jackson, Social and Economic Networks, Princeton University Press, 2008.
- [Smid et al, 2004] Jan Smid, Marek Obitko, Václav Snásel: Semantically Based Knowledge Representation, Communications in Computing, 2004.
- [Кохов и др., 2004] Кохов В.А., Незнанов А.А., Ткаченко С.В. Компьютерные методы анализа сходства графов // Девятая Национальная конференция по искусственному интеллекту с международным участием. КИИ-2004: Труды конференции. В 3-х т. Т. 1. – М. : Физматлит, 2004.
- [Кохов, 2002] Кохов В.А. Концептуальные и математические модели сложности графов. // М. : Издательство МЭИ, 2002.
- [Незнанов и др., 2008] Незнанов А.А., Кохов В.А. Программные средства для построения и исследования моделей структурной сложности и сходства // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием. КИИ-2008. Труды конференции. Т. 1. – М. : ЛЕНАНД, 2008.