УДК 517.977.5

В. М. Дубовой, д. т. н., проф.; А. М. Москвин

РАЗРАБОТКА СИСТЕМЫ НЕЧЕТКОЙ КЛАССИФИКАЦИИ ГИПЕРТЕКСТОВЫХ СТРУКТУР

В статье предложена нечеткая система оценки оптимальности гипертекстовых информационных структур. Исследованы существующие метрики для формализации их функционально значащих параметров на основе данных, полученных в процессе исследования влияния вида гипертекстовых структур на значения показателей их оценки.

***Ключевые слова**: гипертекст, гипертекстовая метрика, индекс компактности, индекс стратификации, нечеткая классификация, оптимальность структуры.*

Введение

Гипертекстовые информационные системы чрезвычайно распространены, поскольку они являются основой Интернет-ресурсов. Сеть гипертекстовых фрагментов характеризуется сложной, хаотической и неоптимизированной структурой, что значительно усложняет поиск необходимой информации. **Проблема** объективного оценивания характеристик гипертекста с целью его оптимизации с развитием Интернет-технологий приобретает все большую **актуальность**.

В теории гипертекста для формализации его функционально значащих параметров существует специальная гипертекстовая метрика [1], которая включает два оценочных показателя – индекс информационной компактности и индекс стратификации. В качестве модели гипертекстовой информационной среды выбран ориентированный граф, в котором вершинами являются соответствующие фрагменты, а ребрами – связи между ними.

Индекс информационной компактности характеризует степень пересечения гипертекстовой структуры связями [2]:

$$Cp = \frac{CD_{\max} - CD}{CD_{\max} - CD_{\min}}, \qquad (1)$$

где $CD_{\max}$ – максимально возможное количество шагов, которые необходимо пройти по ссылкам, связывающим все узлы гипертекста; $CD_{\min}$ – минимально возможное количество шагов, которые связывают все узлы гипертекста; CD – показатель путей в графе, для определения которого необходим предварительный расчет преобразованной матрицы расстояний.

Значение индекса информационной компактности изменяется в пределах [0; 1], что допускает сравнение гипертекстовых документов между собой. Абсолютно несвязанный гипертекст характеризуется нулевым значением индекса информационной компактности – $Cp=0$, и наоборот, абсолютно связанный – значением $Cp=1$. Высокий уровень компактности характеризует такие гипертекстовые структуры, в которых на любой из информационных блоков можно с легкостью попасть из любого другого блока, это, обычно, обеспечивается большим числом перекрестных ссылок. Следует отметить, что слишком высокая компактность может привести к полной дезориентации пользователя гипертекстовой системы. В свою очередь, низкая информационная компактность способствует выпадению из поля зрения многих фрагментов гипертекста или приводит к потере достижимости отдельных фрагментов.

Индекс стратификации детально рассмотрен в [1] и введен для характеристики линейности гипертекста [3]:

$$St = \frac{AP}{LAP}, \tag{2}$$

где AP – абсолютная стратификация, LAP – линейная абсолютная стратификация гипертекста с n узлами идентичная абсолютной стратификации линейного гипертекста аналогичной размерности. Рассчитывается по формуле:

$$LAP = \begin{cases} \frac{n^3}{4}, & \text{если } n \text{ парное} \\ \frac{n^3 - n}{4}, & \text{если } n \text{ непарное.} \end{cases} . \tag{3}$$

В случае абсолютно стратифицированного гипертекста, индекс стратификации принимает значение $St=1$ и, наоборот, – $St=0$. Фактически, индекс стратификации позволяет оценить уровень связанности элементов, стоящих на разных уровнях иерархии.

Доля отсутствующих путей по смыслу связана с индексом информационной компактности:

$$K_m = \frac{Q_m}{n^2 - n}, . \tag{4}$$

где Q_m – количество отсутствующих путей в графе [3]. Для расчета коэффициента отсутствующих путей, необходимым является предварительное определение матрицы расстояний графа.

Максимальное количество отсутствующих путей равно $n^2 - n$, минимальное – 0. Значение доли отсутствующих путей изменяется в пределах [0; 1] и допускает сравнение систем гипертекстовых документов между собой.

Цикломатическое число характеризует отличие структуры графа от деревовидной структуры и определяется по формуле:

$$Cn = m(G) - n(G) + p, \tag{5}$$

где m(G) – число ребер, n(G) – число вершин, p – число связанных компонент графа [3].

Цикломатическое число показывает наименьшее количество ребер, которые необходимо удалить для того, чтобы граф стал деревом. Для сильно связанного графа p=1.

Анализ критериев оценки гипертекстовой структуры показал не универсальность и невозможность их раздельного использования для получения адекватной характеристики структуры в связи с их функциональной ограниченностью.

Примером, иллюстрирующим недостаточность каждого показателя для оценки качества гипертекста, являются результаты, приведенные на рис. 1. Индекс стратификации остается одинаковым как для древовидной структуры без внутри-иерархических связей, так и с ними. С другой стороны, значение индекса компактности, как для линейной замкнутой структуры, так и для иерархической, является почти одинаковым.

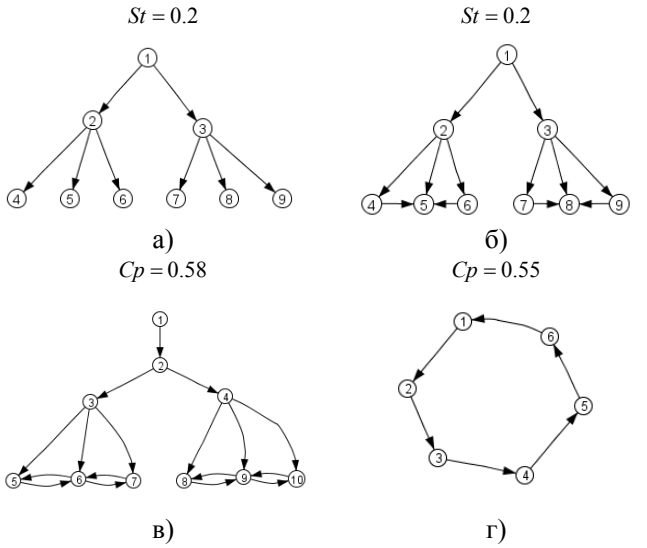


Рис. 1. Исследование зависимости вида структур от значения их показателей а, б – индекса стратификации; в, г – индекса информационной компактности

Задачей работы является построение системы нечеткой оценки качества гипертекстовых структур как один из путей комплексного использования существующих показателей.

Для оценки влияния структуры гипертекста на значения критериев были проведены исследования по видоизменению гипертекстовых иерархических структур и вычислению для них значений рассмотренных ранее показателей.

Базовая иерархическая структура, над которой производились операции модификации, является деревом и содержит только однонаправленные связи. Эта структура характеризуется четкой стратификацией и имеет вид представленный на рис. 2. Исследования проводились итерационным способом, при котором, на каждом шаге, к графу, с помощью разработанного генератора иерархических структур, добавлялось по одному ребру и производилось вычисление значений индекса информационной компактности, индекса стратификации, доли отсутствующих путей и цикломатического числа графа.

Проведено исследование влияния разных типов связей на значения критериев оценки гипертекстовой структуры:

- двусторонних связей между соседними уровнями;
- однонаправленных связей с нижних уровней к уровням более высокой иерархии;
- однонаправленных связей с высших уровней к уровням более низкой иерархии;
- горизонтальных связей между элементами отдельных субиерархий;
- горизонтальных связей между элементами различных субиерархий.

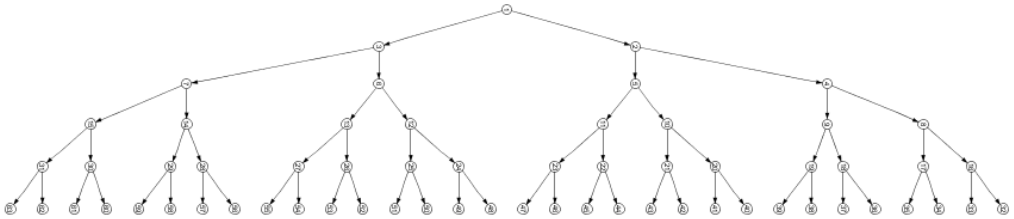


Рис. 2. Вид базовой иерархической структуры

Графически принцип изменения структуры графа использованной в данном исследовании показан на рис. 3. С нижних уровней добавляются обратные связи ко всем вышестоящим вершинам в каждой субиерархии. Эти связи обозначены на рис. 3 штрихпунктирными линиями.

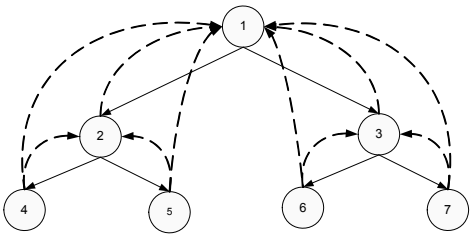


Рис. 3. Графическая интерпретация принципа видоизменения структуры графа

Как видно из результатов исследований (рис. 4), изменение значения индекса стратификации носит нелинейный характер. Начальное увеличение значения индекса стратификации обусловлено появлением связей между конечными (теми, что не имеют исходящих связей) и более высокими в иерархии вершинами. Медленный спад индекса стратификации обусловлен усложнением структуры перекрестными ссылками.

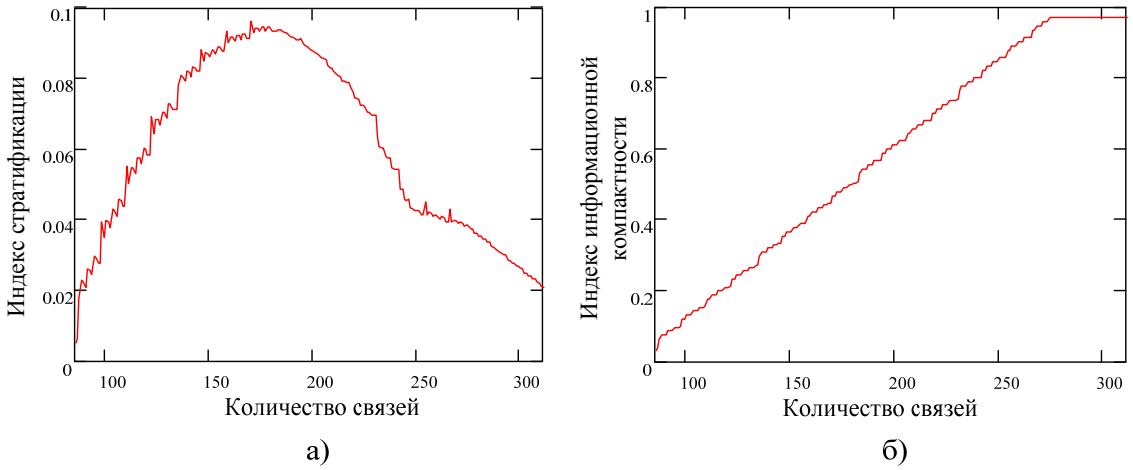


Рис. 4. Графические зависимости изменения критериев оценки гипертекстовой структуры от количества однонаправленных связей типа “снизу-вверх” в графе:
а) индекса стратификации; б) индекса информационной компактности

Проведенные исследования позволяют установить эвристическую связь между изменениями отдельных характеристик гипертекста и его качеством:

1. Изменение гипертекстовой структуры путем добавления обратных связей только на 1 уровень выше дает возможность пользователю возвращаться назад в процессе навигации, но только на один шаг. Данное средство предоставляет возможность сделать гипертекст достижимым из любой вершины. Отдельное использование только данного подхода является неудачным, поскольку предусматривает большое количество переходов пользователя между узлами в процессе поиска информации. Более приемлемым подходом к модификации гипертекстовой структуры является такой, в котором, кроме обратных связей, добавляются однонаправленные связи к уровням более высокой иерархии. Скорость достижимости фрагментов в такой структуре будет выше. Оптимальной сложностью будем считать такую, которая отвечает точке перегиба характеристики зависимости индекса стратификации к количеству связей и которая соответствует началу вырождения структуры.

2. Введение однонаправленных связей с более высоких уровней до нижних уровней иерархии, как показали исследования, разрушает стратификацию и почти не влияет на изменение индекса информационной компактности. Поэтому отдельное использование только данного подхода является неэффективным, т. к. хоть и увеличивается скорость достижимости с верхних уровней на нижние, верхние уровни становятся перегруженными ссылками, не давая возможности пользователям повторного к ним возвращения.

3. Горизонтальные связи между элементами разных субиерархий незначительно влияют на показатель стратификации, хотя серьёзно влияют на индекс информационной компактности. Действительно, скорость достижения фрагментов значительно увеличивается за счет появления связей между разными иерархиями. Оптимальной сложностью такой структуры будем считать такую, которая отвечает точке перегиба характеристики зависимости индекса стратификации к количеству связей.

Обобщим полученные результаты в виде нечеткой системы оценки качества гипертекстовой структуры. Использование нечеткой логики для данной задачи является оправданным, поскольку полученные зависимости вида гипертекстовой структуры от значений ее показателей имеют сложный нелинейный характер. Поиск аналитических зависимостей, которые их описывают, является проблематичным, а оценочные параметры в некоторых случаях могут давать противоречивые результаты.

Исследование и разработка нечеткой системы проводились в пакете прикладных программ Matlab 6, с помощью инструментария Fuzzy Logic Toolbox.

Согласно результатам исследований, была сформирована нечеткая база знаний классификации результатов оценки. Для создания системы нечеткого логического вывода определены 3 лингвистические переменные – индекс информационной компактности, индекс стратификации и доля отсутствующих путей. Для системы нечеткого логического вывода предлагается использование системы типа Мамдани, в которой значения входных и выходной переменных задаются нечеткими термами [4]. База знаний состоит из 19 нечетких правил приведенных в таблице 1.

Таблица 1

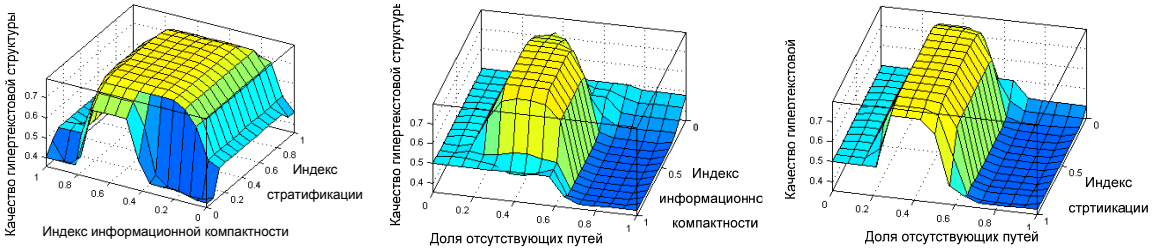
Правила нечеткой базы знаний

Индекс стратификации	Индекс информационной компактности	Доля отсутствующих путей	Качество гипертекстовой структуры
высокий	высокий	высокий	низкое
высокий	ниже среднего	средний	среднее
средний	высокий	низкий	низкое
средний	высокий	низкий	высокое
средний	высокий	средний	низкая
средний	выше среднего	низкий	среднее
средний	выше среднего	средний	среднее
средний	выше среднего	средний	высокое
средний	ниже среднего	низкий	низкое
средний	ниже среднего	средний	среднее
средний	средний	высокий	низкое
средний	средний	низкий	низкое
средний	средний	низкий	высокое
средний	средний	средний	высокое
низкий	высокий	низкий	низкое
низкий	высокий	средний	низкое
низкий	ниже среднего	средний	низкое
низкий	низкий	низкий	низкое
низкий	низкий	средний	низкое

Визуализация поверхности “входы-выходы” произведена с помощью модуля Surface Viewer из пакета прикладных программ Matlab.

На рис. 5 изображены поверхности “входы-выходы” для выходной переменной от комбинации входных переменных – индекс стратификации, индекс информационной компактности и доля отсутствующих путей. Согласно полученным результатам, терму «высокий» функции принадлежности выходной переменной отвечает определенная выпуклая область.

Визуализации, представленные на рис. 5, подтверждают, что область оптимальных решений представляет собой определенное их множество. Диапазоны изменения значений параметров оценки для области оптимальных решений согласно результатам представленными на рис. 5 (а, б, в) являются следующими - [0.2; 0.85] для индекса информационной компактности, [0.1; 0.8] для индекса стратификации и [0.25; 0.6] для доли отсутствующих путей.



а) б) в)

Рис. 5. Поверхность “входы-выход” в SurfaceViewer: а) входы – индекс информационной компактности, индекс стратификации; б) входы – доля отсутствующих путей, индекс стратификации; в) входы – доля отсутствующих путей, индекс информационной компактности

Вывод. Разработанная система нечеткой классификации семантической структуры

Интернет-ресурсов на основе 3-х показателей (индекса информационной компактности, индекса стратификации и доли отсутствующих путей) позволяет учесть особенности каждого показателя при классификации структуры. Разработанные правила могут быть использованы для построения автоматизированной системы оценки семантической структуры сайтов с помощью инструментария FuzzyJ, который представляет собой набор библиотек реализующих механизмы нечеткой логики для языка Java.

СПИСОК ЛИТЕРАТУРЫ

1. The Semantic Web [Електронний ресурс] / Tim Berners-Lee, James Hendler, Ora Lassila // Scientific American Magazine. – May, 2001. – Режим доступу до журн.: http://www.sciam.com/article.cfm?id=00048144-10D2-1C70-84A9809EC588EF21.

2. Botafogo R. A. Identifying hierarchies and useful metrics /E. Rivlin, B. Shneiderman // ACM Transactions on Information Systems (TOIS). – 1992. – №2. – P.142 – 180.

3. Harary F. Structural models. An Introduction to the Theory of Directed Graphs / Harary F., Norman R., Cartwright D. – Wiley: New York, 1965. – 415 p.

4. Ротштейн О. П. Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети / Ротштейн О. П. – Вінниця: «УНІВЕРСУМ-Вінниця», 1999. – 320 с.

Дубовой Владимир Михайлович – д. т. н., профессор, заведующий кафедры компьютерных систем управления. тел.: (0432) 598-157, E-Mail: dub@faksu.vstu.vinnica.ua.

Москвин Алексей Михайлович – аспирант кафедры компьютерных систем управления.
Винницкий национальный технический университет.