**А.В. Ганичева, А.В. Ганичев**
*Тверская государственная сельскохозяйственная академия, г. Тверь*
*Тверской государственный технический университет, г. Тверь*

# ГРАФОВЫЙ МЕТОД АНАЛИЗА ТЕКСТОВ
# GRAPH METHOD OF TEXT ANALYSIS

*Ключевые слова: научный текст, структура текста, информационный граф, дерево, вершины, дуги (ребра), сегмент текста, матрица смежности, контур, петля, путь в графе*

*Keywords: scientific text, structure of text, information graph, tree, point or node, oriented edge, text fragment, adjacency matrix, directed circuit, loop, graph path*

## Введение

Автоматическое распознавание текста является одним из наиболее быстро развивающихся направлений искусственного интеллекта. Возможные направления использования систем автоматического анализа текстов для решения различных задач приведены в статье (Ганичева, 2016).

В системах автоматического анализа текстов часто применяются методы теории графов. Привлечение графовых моделей в системах автоматического анализа текстов вызвано необходимостью такой формы описания текста, которая была бы компактной, строго формализованной, наглядной. Перечислим некоторые примеры применения графовых моделей. В статье (Голубев, 2011) для распознавания соответствия документа эталонному документу производится сравнение модельного графа и графа, полученного по изображению документа. В статье (Целых, 2008) предлагается использовать нечеткий граф для изучения коммуникаций в социальной среде. Графовая модель используется в работе (Карпенко 2011) для оценки релевантности проверяемого документа образцам из онтологической базы знаний. В статье (Тревгода, 2009) графовая модель используется для реферирования текста.

## Особенности анализа структуры научных статей

Научный текст имеет строгую внутреннюю организацию составляющих его логико-смысловых частей. Текст можно считать состоящим из отдельных сегментов, между которыми установлены связи (отношения). В графовой модели сегменты можно считать вершинами графа, а связи между сегментами – ребрами (дугами) графа. Поэтому текст можно представить соответствующим графом, что позволяет формализовать процесс анализа текста и использовать хорошо развитый математический аппарат теории графов.

Под структурой текста понимается его внутренняя организация. Единицами внутренней структуры текста являются: набор слов, предложение или совокупность предложений, логически объединенных в единый сегмент.

Единицы текста находятся в логической взаимосвязи и связаны различными отношениями: подчинения, несовместимости, наследования и т.д.

Структура научных статей обычно состоит из восьми частей: заголовок, перечень авторов, ключевые слова, (аннотация) реферат, введение, основная часть, заключение, список литературы.

Во введении можно выделить следующие части: актуальность проблемы, известные варианты решения проблемы (аналоги), достоинства и недостатки известных методов, цель и задачи исследования.

Основная часть может содержать: описание предлагаемого варианта решения проблемы, оценку новизны предлагаемого метода, определение места исследования в системе знаний, перечень используемых технических средств и оборудования, метод исследования, результаты экспериментальной проверки разработанных методов.

Заключение может включать: перечень полученных результатов, выводы, преимущества предложенного варианта решения проблемы по сравнению с аналогами, рекомендации по использованию и применению разработанного метода.

Отдельные части научной статьи, а также составляющие этих частей будем называть сегментами текста.

## Анализ информационного графа по матрице смежности

Для исследования передачи и преобразования информации в сложных системах применяются информационные графы (Ганичева, 2005). Вершины информационного графа соответствуют этапам или ключевым моментам обработки информации (например, исходные или начальные данные, промежуточные и окончательные результаты обработки информации). Дуги графа указывают на порядок взаимодействия вершин в направлении передачи и преобразования информации. Матрицы смежности информационных графов используются при анализе схем потоков информации в сложных системах.

Рассмотрим применение информационного графа для анализа текста на простом примере. Пусть текст содержит следующие сегменты; основная цель текста; задачи, раскрывающие основную цель; основная часть текста; заключение.

Такую структуру текста можно представить информационным графом, изображенным на рис. 1.

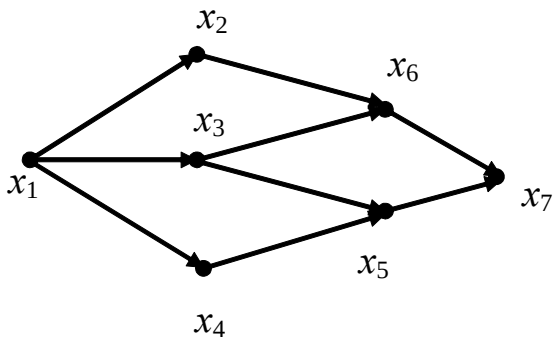


Рис. 1. Информационный граф текста.

На рис. 1 вершина $x_1$ соответствует основной цели текста, $x_2$, $x_3$, $x_4$ – задачам, раскрывающим основную цель, $x_5$, $x_6$ – сегментам основной части текста, $x_7$ – заключению.

Между вершинами информационного графа существует отношение порядка, оно разбивает весь процесс движения информации от начала до заключения на такты (этапы), в результате которых формируются вершины графа. В информационном графе для формирования любой вершины $x_i$ необходимо, чтобы информация поступала в эту вершину по всем путям, ведущим из исходных данных в $x_i$.

Порядком вершины $x_i$ информационного графа является число, равное максимальной из длин путей, ведущих в вершину $x_i$ из начальной вершины. Так, в рассмотренном примере порядок вершин $x_2$, $x_3$, $x_4$ равен 1, порядок $x_5$, $x_6$ равен 2. В этом случае говорят, что вершины $x_2$, $x_3$, $x_4$ формируются в результате первого такта, а $x_5$, $x_6$ – на втором такте. Порядком информационного графа называется максимальное числи тактов обработки информации, необходимое для достижения конечного результата. Порядок информационного графа равен наивысшему из порядков вершин, отвечающих окончательным результатам. В примере информационный граф имеет порядок 3, так как максимальное число тактов в движении информации, необходимое для получения заключения $x_7$, равно 3.

Построим матрицу смежности данного графа и найдем 2-ю, 3-ю и 4-ю степень матрицы смежности:

$$A=\begin{pmatrix}0&1&1&1&0&0&0\\0&0&0&0&1&0&0\\0&0&0&0&1&1&0\\0&0&0&0&0&1&0\\0&0&0&0&0&0&1\\0&0&0&0&0&0&1\\0&0&0&0&0&0&0\end{pmatrix},\ A^2=\begin{pmatrix}0&0&0&0&2&2&0\\0&0&0&0&0&0&1\\0&0&0&0&0&0&2\\0&0&0&0&0&0&1\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{pmatrix},\ A^3=\begin{pmatrix}0&0&0&0&0&0&4\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{pmatrix},$$

$$A^4=0.$$

Рассмотрим, какую информацию о структуре текста можно получить из анализа этой матрицы смежности и ее степеней.

Если *j*-ый столбец состоит из одних нулей, то вершина $x_j$ – начальные данные (основная цель текста в примере), если *i*-ая строка состоит из одних нулей, то $x_i$ – заключение.

Порядок вершины $x_j$ равен наивысшему показателю $k$ ($k$ меньше порядка матрицы) такому, что в матрице $A^k = A \cdot A \cdot \ldots \cdot A$ (справа стоит $k$ сомножителей $A$) в столбце с номером $j$ имеется хотя бы один отличный от нуля элемент. Порядок вершины позволяет определить номер того такта (этапа), после которого тот или иной сегмент текста перестает учитываться в последующей обработке информации. А именно: номер такта, после которого сегмент $x_i$ может не учитываться при анализе текста, равен максимальному из порядков вершин, отвечающих отличным от нуля элементам *i*-ой строки матрицы смежности $A$.

Матрицы $A \neq 0,\ A^2 \neq 0,\ A^3 \neq 0,$ а матрица $A^4 = 0.$ Следовательно порядок графа равен 3, что согласуется с полученным ранее результатом.

В рассматриваемом примере отличным от нуля элементам столбцов матрицы $A^2$ отвечают вершины $x_5$, $x_6$. Следовательно, порядки этих вершин равны 2. Таким образом, после второго такта задача $x_3$ не анализируется в тексте. Порядок вершины $x_7$ будет равен 3, т.к. в матрице $A^3$ в 7-м столбце есть отличный от 0 элемент.

Для дальнейшего анализа строится матрица $B = A + A^2 + \ldots + A^n$ (где $n$– порядок $A$):

$$B = \begin{pmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 4 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Элементы этой матрицы указывают число возможных путей в информационном графе. Элемент $b_{17} = 4$ указывает, что в графе имеется 4 пути, ведущих из $x_1$ в $x_7$; из $x_3$ в $x_7$ ведут два пути, т.к. $b_{37} = 2$; а из $x_1$ в $x_5$ и $x_6$ ведут два пути, т.к. $b_{16} = b_{15} = 2$; из $x_2$ в $x_6$ путей нет, поскольку $b_{26} = 0$ и т.д. Отличные от нуля элементы, стоящие в *j*–ом столбце матрицы $B$, указывают на результаты, участвующие в формировании результата $x_j$, а именно: порядковые номера отличных от нуля элементов *j*–ого столбца равны номерам результатов, из которых формируется результат $x_j$. Для рассмотренного примера в формировании, например, $x_2$ участвует $x_1$; в формировании $x_7$ участвуют $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$ и т.д. На практике эти сведения используют, когда обнаружено нарушение логической связи в получении некоторого сегмента.

Отличные от нуля элементы матрицы $B$, стоящие в *i*–ой строке, перечисляют все результаты, при формировании которых использовался результат $x_i$, а именно: номера отличных от нуля элементов *i*–ой строки

равны номерам результатов, в формировании которых участвовал результат $x_i$. В примере в формировании результатов $x_5$, $x_6$, $x_7$ участвовал результат $x_3$, а в формировании $x_7$ – только сегмент $x_5$.

Описанную методику анализа информационного графа целесообразно использовать для распознавания структуры текста.

**Графовый метод определения структурированности текстов**

Основными конструктивными признаками текста являются целостность и связность. Они отражают содержательную и структурную сущность текста. Рассмотрим как на основе матрицы смежности графа определить целостность и связность текста.

При анализе текстов особое значение имеет выделение сегментов, соответствующих изолированным, висячим и тупиковым вершинам графа. Изолированные вершины не инцидентны ни одному из ребер (дуг) графа, что может свидетельствовать о том, что данный сегмент графа не связан с другими сегментами. Висячие вершины соответствуют сегментам, в которые нельзя попасть из других сегментов. Тупиковые вершины показывают, что из данных сегментов нельзя попасть в другие сегменты текста.

Отыскать на графе изолированные, висячие и тупиковые вершины можно по матрице смежности графа A=$\|a_{ij}\|$, по которой для каждой вершины $k$ ($k=\overline{1,n}$, $n$ – число вершин в графе) определяется вектор $a(k)=(a_k, a^k)$ с компонентами:

$a_k = \sum_{j=1}^{n} a_{kj}$, $a^k = \sum_{i=1}^{n} a_{ij}$, где $a_k$ - сумма элементов $k$-ой строки, $a^k$ -$k$-го столбца матрицы смежности.

Величина $a_k$ определяет число дуг, выходящих из вершины $k$, а $a^k$ - число дуг, входящих в нее. Когда $a_k = a^k = 0$, вершина $k$ будет изолированной, если $a_k = 0$ – тупиковый, а при $a^k = 0$ – висячей.

Наличие в графе изолированных вершин обычно свидетельствует о не связанности (отсутствии целостности) текста.

Висячие вершины должны соответствовать заключительным положениям текста, а тупиковой вершиной может быть только сегмент, соответствующей центральной идее текста.

В приведенном примере информационного графа текста на основе анализа матрицы смежности можно сделать вывод, что в данном графе нет изолированных, тупиковых и висячих вершин.

Исследование особенностей связей между сегментами текста направлено, прежде всего, на выявление в соответствующем графе петель, контуров и сильно связанных подграфов. Петля интерпретируется как наличие связи между входом и выходом одного и того же сегмента, т.е. замкнутость рассуждений в данном сегменте. Контур образует путь – чередующуюся последовательность ребер (дуг) и вершин, в котором начальная и конечная вершина совпадают. Данное обстоятельство говорит о

возврате к одним и тем же рассуждениям, т.е. отсутствию причинно – следственных связей в тексте.

Подграф является сильно связным, если все входящие в него вершины достижимы, когда из любой вершины подграфа можно попасть в любую другую его вершину, т.е. все сегменты текста достижимы из других сегментов. Такая структура не свойственна научным текстам.

Наличие петли, контура и сильно связанных подграфов так же возможно определить на основе матрицы смежности. Так о наличии петли будет свидетельствовать ненулевой элемент матрицы смежности. О наличии контуров свидетельствует равенство $a_{ij}=a_{ji}=1$. Главный определитель матрицы $\|a_{ij}\|$ характеризует число замкнутых циклов так, что каждое его слагаемое за исключением диагональных соответствует одному их циклов. Слагаемые диагонального минора $M_{ij}^{\,n-1}$ матрицы $\|a_{ij}\|$ характеризуют число и характер замкнутых циклов, остающихся в структуре после исключения $i$-го элемента.

Для анализа связности (целостности) текста понятие связности графа мало подходит. Удобнее для анализа связности графа использовать показатель $\alpha$, характеризующий относительную разность числа связей $R$, имеющихся в данном тексте, и числа связей $R_{min}$, минимально необходимого для связности (целостности) графа текста. Показатель $\alpha$ может интерпретироваться как мера избыточности текста по связям (ссылкам).

Если граф содержит $n$ – вершин, то $R_{min}=n-1$ независимо от того, является граф ориентированным или нет (т.е. граф имеет древовидную структуру). Следовательно, $\alpha=(R-R_{min})/R_{min}= R/(n-1)-1$.

Значение $R$ определяется по матрице смежности для ориентированных и неориентированных графов по-разному. В ориентированном графе каждой дуге $(i,j)$ соответствует единственный элемент матрицы смежности $a_{ij}=1$, а в неориентированном графе таких элементов будет два: $a_{ij}= a_{ji}=1$. Поэтому для ориентированного и неориентированного графов, соответственно, имеем:

$$R=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij};\;\; R=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}.$$

Для не избыточных текстов значение $\alpha$ должно быть мало.

Для рассматриваемого в примере ориентированного графа $R=9$, $n=7$, поэтому $\alpha=0,5$. Таким образом, структура текста в рассмотренном примере является не избыточной (имеется только одна лишняя связь).

## Заключение

Описание и использование информационного графа с помощью матрицы смежности упрощает анализ текста, а в ряде случаев является пока единственно возможным методом такого анализа.

## Литература

1. Ганичева А.В., Ганичев А.В. Дискурсный метод распознавания структурированности текстов // Мир лингвистики и коммуникации: электронный научный журнал. - № 2, 2016. – С. 31 – 38. - Режим доступа: tverlingua.ru

2. Голубев С.В. Распознавание структурированных документов на основе машинного обучения // Бизнес-информатика. - № 2 (16), 2011. – С. 48 – 55.

3. Целых Ю.А. Теоретико-графовые методы анализа нечетких социальных сетей // Программные продукты и системы. - № 2, 2008. – С. 48 – 50.

4. Карпенко А.П. Методика оценки релевантности документов онтологической базы знаний // Информационные технологии. - № 4,2011. – С. 13 - 23.

5. Тревгода С.А. Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений: автореф. дис. ... канд. техн. наук. – СПб., 2009. – 18 с.

6. Ганичева, А.В. Математика для психологов / А.В. Ганичева, В.П. Козлов. – М.: Аспект-Пресс, 2005. – 239 с.

**(0,4 п.л.)**