Мартинович Е.Н.
Шуткин А.С.
Попов Е.В.
Научный руководитель: кандидат физико-математических наук
Семенов М.Е.

АВТОМАТИЗАЦИЯ РАБОТЫ, С ПОМОЩЬЮ ПРИМЕНЕНИЯ XML ГРАФОВ В ВЫЯВЛЕНИИ АНОМАЛИЙ ПРИ АНАЛИЗЕ ДОКУМЕНТОВ

С целью автоматизации работы аудиторских компаний, минимизации человеческого фактора и уменьшения затрат ресурсов, нами был придуман и реализован алгоритм XML графов для анализа документов. Из анализа литературы мы увидели, что таких работ еще не было представлено – полностью инновационная разработка.

В качестве входных данных мы используем подмножество языка разметки XML, потому что отчетность чаще всего представляется именно в таком виде. Сначала производится анализ XML-файлов и составление списков и словарей содержащихся в них тегов, их значений и иерархии. На основании этого составляется текстовые документ с подробным перечнем связей по каждому отдельному элементу.

Далее строится граф первого вида, изображающий связи файлов через отдельные теги, каждый элемент имеет всплывающую подсказку с информацией о расположении данного элемента в текстовом файле. Вершинами графа считаются элементы разметки файла, а ребрами – связь между ними. На рисунке 1 мы можем увидеть, как соединяются иерархии двух файлов. Например, тег со значением «дерево» будет соответствовать такому же в другом файле, и можно установить главную тематику документа. Также мы выявили аномалию слово «компьютер», которое сложно отнести к какой-либо группе данных файлов.

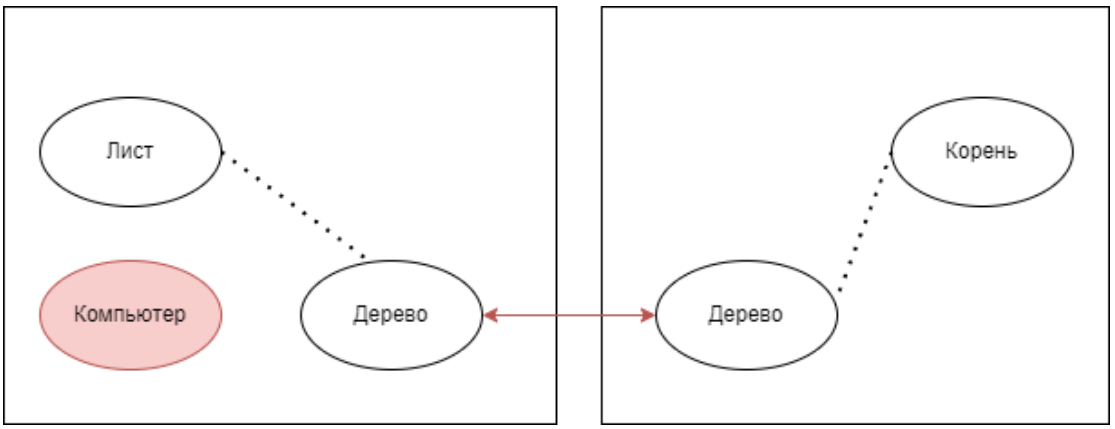

Рисунок 1 – Пример соединения файлов

На основании выстроенной иерархии производится подсчет весов тегов и их значений, сил взаимодействия файлов друг с другом, создание таблицы связей файлов по полям и значениям. Создается и нормализуется таблица величины связей между файлами, на основании которой производится понижение размерности с помощью метода главных компонент. Полученная кластеризация диагонализируется и высчитывается удаленность первоначального положения от итогового. Затем файлы распределяются по холсту в соответствии с их изначальным положением на диагонали и углом, пропорциональным вычисленной удаленности. В итоге строится граф второго вида визуализируется в трех вариантах: с низкой, средней и высокой степенями детализации.

В результате применения нашей разработки мы получаем:

- relations.txt – текстовый файл, с подробным перечнем связей по каждому отдельному элементу.
- result.csv – таблица связей объектов графа через поля.
- test.svg – граф первого типа. Показывает взаимосвязи XML файлов через общие теги внутри файла и положения описанных элементов в текстовом документе.
- low.png – граф второго типа с низкой степенью детализации, показывающий взаимосвязи XML-файлов друг с другом.
- middle.png – граф второго типа с средней степенью детализации, показывающий взаимосвязи XML-файлов друг с другом.
- high.png – граф второго типа с высокой степенью детализации, показывающий взаимосвязи XML-файлов друг с другом.