УДК 004.6

В.А. Губин, Ю.Ю. Шевякова

*Харьковский национальный университет радиоэлектроники, Харьков*

# КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ФРАГМЕНТОВ СЛАБОСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ КАК АТРИБУТ ДАННЫХ ИЛИ КАК ЗНАЧЕНИЕ АТРИБУТА ДАННЫХ

*Предложен метод, позволяющий отнести тот или иной обособленный текстовый фрагмент слабоструктурированного текстового документа к атрибуту данных или к значению атрибута данных. При этом предполагается, что анализируется совокупность слабоструктурированных текстовых документов одного вида. Рассматриваемый подход основан на сопоставлении абсолютных частот встречаемости текстовых фрагментов в совокупности документов.*

***Ключевые слова:*** *слабоструктурированный документ, данные, атрибут данных, значение данных.*

## Введение

**Постановка задачи.** Объектом исследования в данной работе является совокупность слабоструктурированных текстовых документов. В статье [1] приведены примеры классов такого рода документов, представлены их основные признаки. В частности, выделяются следующие их признаки и свойства как источников данных: документ содержит внутреннюю разметку; содержимое документа разбито внутренним форматированием на обособленные текстовые фрагменты; каждый фрагмент объективно представляет собой либо атрибут данных, либо значение атрибута данных, во внутренней разметке документов нет формальных признаков, указывающих на то, что есть атрибут данных, а что есть значение атрибута данных

В данной статье решается задача классификации обособленных текстовых фрагментов слабоструктурированных текстовых документов. Это создаст предпосылки к формализации данных, содержащихся в таких документах. По сути, необходимо отнести каждый такой фрагмент либо к атрибуту данных, либо к значению атрибута данных. В основе подхода, рассматриваемого в данной работе, лежит то, что обрабатывается совокупность документов как единое целое.

Предполагается, что каждый слабоструктурированный текстовый документ представлен в объектном виде. В работе [2] показано, что каждый такой документ в этом случае представлен в виде совокупности объектов двух видов. Одна совокупность – это объекты контейнеры, отражающие структуру документа. Другая совокупность – атомарные объекты. При этом каждый атомарный объект соответствует некоторому обособленному текстовому фрагменту документа.

**Анализ последних достижений и публикаций.** Существует ряд работ, посвященных решению задачи доступа к данным, содержащимся в текстовых документах, извлечению информации из такого рода документов, идентификации пар «атрибут-значение». Так, в работе [3] для решения задачи извлечения информации из неструктурированных источников, таких как документы и электронные письма, предлагается разработка и использование контекстно-свободных грамматик с последующим их обучением на выборке данных. В работе [4] предлагается метод идентификации пар «атрибут-значение», основанный на предварительном задании определенных типов структурных контекстов и фильтров, являющихся критериями отбора текстовых строк, предположительно являющихся парами «атрибут-значение». В [5] изначально определяются классы, содержащие описание атрибутов данных и их возможные значения. Затем осуществляется отбор из сети документов, содержащих определенные в классе атрибуты с последующим возможным доступом к их значениям.

Общим для такого рода работ является то, что документы рассматриваются изолированно от других документов того же типа. Недостатком такого подхода является то, что необходимо выполнять предварительную, порой достаточно серьезную, работу по заданию критериев отбора пар «атрибут-значение».

В данной работе рассматривается совокупность однотипных документов как единое целое. При этом нет необходимости в том, чтобы все эти документы имели один тип. В рассматриваемой совокупности могут быть несколько групп однотипных документов. Для получения позитивных результатов достаточно того, чтобы каждый тип документа был представлен несколькими документами. Это дает возможность в качестве критерия того, чем является тот или иной текстовый фрагмент документа, взять частоту встречаемости соответствующей текстовой строки в совокупности однотипных документов. Данный критерий

является достаточно универсальным и, таким образом, позволяет избавиться от необходимости разрабатывать специальные критерии для каждого сценария появления данных в документе.

**Цели и задачи исследования.** Целью данной работы является решение задачи классификации обособленных текстовых фрагментов слабоструктурированных текстовых документов как атрибутов данных, либо как значений атрибута данных.

Для решения этой задачи необходимо предварительно каждый документ из исследуемого множества документов представить в виде совокупности атомарных объектов и объектов контейнеров.

Абсолютная встречаемость атомарных объектов в документах по ряду причин не может служить надежным признаком того, что они являются либо атрибутом, либо значением данных. Более информативным будет сопоставление абсолютной встречаемости атомарных объектов, находящихся в отношении ассоциации. Как отмечается в работе [2], отношение ассоциации между атомарными объектами отражает предположение о том, что соответствующие им текстовые фрагменты документа образуют пару «атрибут-значение».

В ходе сопоставления частот абсолютной встречаемости получаются оценки вероятности того, что тот или иной атомарный объект является атрибутом данных или значением атрибута данных. Если оценки вероятностей равны нулю или единице, то в этом случае имеются все основания сделать соответствующий однозначный вывод. В случае получения промежуточных результатов необходимо выдвинуть и проверить статистическую гипотезу о значении вероятности того, чем является атомарный объект. По итогам этой проверки неоднозначность должна быть либо устранена, либо делается вывод о том, что текстовые фрагменты документа, соответствующие некоторым атомарным объектам, в силу ряда причин, невозможно однозначно классифицировать как атрибут данных, либо как значение атрибута данных.

## Подготовительная работа

Рассмотрим совокупность электронных текстовых документов $\Omega$, содержащую N документов $D_1$, $D_2$, … , $D_N$. Таким образом:

$$\Omega = \{D_1, D_2, ..., D_N\}.$$

Обозначим через $\Psi$ множество всех атомарных объектов, содержащихся в анализируемой совокупности текстовых документов. Множество $\Psi$ в этом случае можно представить таким образом:

$$\Psi = \{\Psi_{D_1}, \Psi_{D_2}, ..., \Psi_{D_N}\},$$

где $\Psi_{D_i}$ – множество атомарных объектов, содержащихся в i-м документе. Учитывая, что каждый атомарный объект в качестве одного из своих свойств имеет ссылку на документ, в который он входит, то множество $\Psi$ можно представить в виде сплошной совокупности атомарных объектов:

$$\Psi = \{\psi_1, \psi_2, ..., \psi_{N\psi}\},$$

где $N_\psi$ – количество атомарных объектов в совокупности документов $\Omega$.

Ответ на вопрос, чем же является текстовая строка, соответствующая некоторому атомарному объекту – атрибутом данных или значением атрибута данных, можно попытаться дать, подсчитав частоту встречаемости этой текстовой строки в совокупности документов. В первом приближении можно считать, что чем чаще встречается в документах одна и та же текстовая строка, тем больше оснований полагать, что она соответствует атрибуту данных и, соответственно, чем реже это происходит, тем больше оснований полагать, что она соответствует значению данных. Но совпадающие тестовые строки могут объективно соответствовать различным данным. Для того, чтобы минимизировать такого рода ошибки, обрабатываются атомарные объекты, каждый из которых соответствует некоторой текстовой строке в документе. Особенностью атомарных объектов является то, что в них, помимо свойства «Текст», есть и другие свойства, в которых отражается контекст появления в документе соответствующей текстовой строки. Это создаст предпосылки для того, чтобы рассматривать текстовые строки с совпадающим значением, но относящиеся к различным данным, как различные.

Таким образом, равенство у атомарных объектов свойства «Текст» автоматически не означает, что они являются частью данных одного и того же типа. Окончательный ответ на вопрос, совпадают или не совпадают конкретные пары атомарных объекты, должна ответить процедура разбиения множества атомарных объектов на классы эквивалентности.

Для решения этой задачи введем на множестве атомарных объектов $\Psi$ отношение эквивалентности E. Важнейшее значение эквивалентности состоит в том, что это отношение определяет признак, который допускает разбиение некоторого множества на непересекающиеся подмножества, называемые классами эквивалентности. И наоборот, всякое разбиение множества на непересекающиеся подмножества определяет между элементами этого множества некоторое отношение эквивалентности [6]. Таким образом, для того, чтобы разбить множество $\Psi$ на классы эквивалентности, необходимо разбить его на непересекающиеся подмножества.

Эта процедура может выглядеть следующим образом. Как уже отмечалось выше, каждый атомарный объект представляет обособленный текстовый фрагмент документа и, наряду с другими свойствами, одним из свойств атомарного объекта есть свойство «Текст» – значение этой текстовой строки. Здесь возможны следующие варианты развития событий:

• У атомарных объектов совпадают все свойства, и они могут считаться тождественными. Такие объекты, безусловно, должны быть отнесены к одному и тому же классу эквивалентности.

• У атомарных объектов соответствующие им текстовые строки различны – эти объекты должны быть признаны безусловно различными не смотря на то, что значения некоторых, а может быть и всех остальных свойств, могут совпадать. Такие объекты должны быть отнесены к разным классам эквивалентности.

• У атомарных объектов совпадают соответствующие им текстовые строки, но есть незначительные отличия в значении других свойств (размер шрифта, например). В этом случае можно как признать так и не признать эти атомарные объекты совпадающими. Они признаются эквивалентными и относятся к одному классу эквивалентности, если эти отличия не превосходят некоторый, установленный предварительно, порог.

Таким образом, в ходе этой процедуры, свойство «Текст» и бинарные свойства являются строго классифицирующими. Свойства, допускающие определенную плавность изменения своих значений, могут отличаться у атомарных объектов из одного класса эквивалентности, если при этом не превзойден допустимый порог отличий. Определив таким образом отношение эквивалентности E на множестве атомарных объектов Ψ, можно представить данное множество в виде совокупности M непересекающихся подмножеств $\bar{\Psi}_i$, каждое из которых представляет некоторый класс эквивалентности:

$$\Psi / E = \{\bar{\Psi}_1, \bar{\Psi}_2, ..., \bar{\Psi}_M\}.$$

Отношение эквивалентности E на множестве атомарных объектов Ψ определило на нем обобщенную форму равенства атомарных объектов. Таким образом, разбиение множества Ψ на классы эквивалентности означает получение эквивалентных между собой атомарных объектов. Т.е. полученные классы эквивалентности $\bar{\Psi}_i$ состоят из тех атомарных объектов, которые рассматриваются в дальнейшем как совпадающие.

## Классификация текстовых фрагментов

Для того чтобы ответить на вопрос, чем же является тот или иной атомарный объект – атрибутом данных или значением атрибута данных, кажется достаточным, на первый взгляд, оценить мощность класса эквивалентности, в который входит данный атомарный объект. Другими словами, подсчитать сколько раз соответствующая данному атомарному объекту текстовая строка встречается в рассматриваемой совокупности текстовых документов. Напрашивается вывод о том, что если это происходит достаточно часто, то данный атомарный объект является атрибутом данных, а если это происходит достаточно редко, то соответствующий атомарный объект являться значением атрибута данных. Но на практике такой подход не срабатывает по двум причинам: во-первых, трудно определиться с тем, что есть «достаточно часто» и что есть «достаточно редко»; во-вторых, некоторые текстовые фрагменты документов, объективно соответствующие значениям данных, вопреки ожиданиям, могут встречаться много чаще, чем текстовые фрагменты, объективно соответствующие атрибутам данных. Решением этой проблемы является учет не абсолютной частоты встречаемости атомарного объекта, а его относительной частоты встречаемости. Для этой цели сопоставляется мощность класса эквивалентности, соответствующего некоторому атомарному объекту и мощности классов эквивалентности, соответствующих атомарным объектам, находящихся с рассматриваемым атомарным объектом в отношении ассоциации.

Отношение ассоциации отражает предположение о том, что атомарные объекты соответствуют текстовым фрагментам документа, образующим пару «атрибут-значение». Ситуации, в которых между двумя текстовыми фрагментами документа может быть установлено отношение ассоциации, рассмотрены в [2]. Данное отношение A можно представить в виде множества упорядоченных пар атомарных объектов:

$$A = \{< \psi_i, \psi_j >, \psi_i \in \Psi, \psi_j \in \Psi\}. \qquad (1)$$

Необходимо заметить, что из того, что $< \psi_i, \psi_j > \in A$ не всегда следует, что $< \psi_j, \psi_i > \in A$ и атомарный объект не может находиться в отношении ассоциации с самим собой.

Далее, для простоты изложения, будем оперировать не классами эквивалентности, а множеством атомарных объектов, каждый из которых представляет один из классов эквивалентности. Т.е. будем оперировать множеством $\bar{\Psi}$ представителей классов эквивалентности:

$$\bar{\Psi} = \{\bar{\psi}_1, \bar{\psi}_2, ..., \bar{\psi}_M\}.$$

При этом для каждого такого атомарного объекта $\bar{\psi}_i$ достаточно сохранять информацию о значении соответствующей текстовой строки, о мощности соответствующего класса эквивалентности и о том, из какого класса эквивалентности он был взят. Через $|\bar{\psi}_i|$ будем обозначать мощность соответствующего i-го класса эквивалентности $\bar{\Psi}_i$, представителем которого является атомарный объект $\bar{\psi}_i$.

Для последующего учета относительной частоты встречаемости атомарных объектов необходимо в отношении ассоциации A перейти от атомарных объектов $\psi_i$ к соответствующим им представителям классов эквивалентности $\bar{\Psi}_i$. В этом случае (1) необходимо переписать так:

$$A = \{< \bar{\psi}_i, \bar{\psi}_j >, \bar{\psi}_i \in \bar{\Psi}_i, \bar{\psi}_j \in \bar{\Psi}_j\}.$$

Необходимо заметить, что количество элементов в множестве A не изменилось.

Объективно, в каждом конкретном случае, атомарный объект представляет либо атрибут данных либо значение атрибута данных. Но до исследований остается неопределенным, чем же на самом де-

ле является текстовая строка соответствующая некоторому атомарному объекту. Т.е. с каждым атомарным объектом можно связать случайную величину «Чем является данный атомарный объект». Эта случайная величина может принимать два значения: «атрибут данных» или «значение атрибута данных» с той или иной вероятностью.

Нахождение оценки этой вероятности осуществляется таким образом: находится отношение числа отношений ассоциации, в которых фигурирует данный атомарный объект предположительно как атрибут данных к общему числу отношений ассоциации, в которых он задействован.

Решение о том, чем является атомарный объект в том или ином отношении ассоциации, принимается на основе сопоставления мощностей соответствующих этим объектам классов эквивалентности. Если для некоторого атомарного объекта мощность класса эквивалентности больше, то данный атомарный объект в этом отношении ассоциации выступает как атрибут данных, а тот атомарный объект, с которым он находится в отношении ассоциации, выступает в этом случае как значение атрибута данных. Если же имеет место совпадение мощностей, то ситуация считается неопределенной и каждому варианту приписывается по 0,5.

Таким образом, для некоторого атомарного объекта $\bar{\psi}_i$ оценка $p_i^*$ вероятности того, что он является атрибутом данных, может быть найдена по такой формуле

$$p_i^* = m_i / n_i,$$

где $n_i$ – количество отношений ассоциации, в которых фигурирует атомарный объект $\bar{\psi}_i$, а $m_i$ – количество отношений ассоциации, в которых $\bar{\psi}_i$ выступает как атрибут данных.

В большинстве случаев будут получаться оценки вероятностей равные единице или нулю. Если $p_i^* = 1$, то текстовый фрагмент, соответствующий атомарному объекту $\bar{\psi}_i$, будет идентифицирован как атрибут данных с вероятностью $p_i = 1$. Если $p_i^* = 0$, то текстовый фрагмент, соответствующий атомарному объекту $\bar{\psi}_i$, будет идентифицирован как атрибут данных с вероятностью $p_i = 0$, т.е. он в этом случае будет идентифицирован как значение атрибута данных.

Случай, когда получена оценка вероятности $p_i^*$ со значением в пределах от нуля до единицы, требует дальнейшего рассмотрения. Такая ситуация возможна в одном из следующих случаев:

- текстовый фрагмент в документе объективно одновременно является атрибутом данных и значением атрибута данных,
- в один класс эквивалентности попали атомарные объекты, объективно относящиеся к различным данным,
- в некоторой паре «атрибут-значение» от документа к документу текстовая строка, объективно соответствующая значению данных, принимает одни и те же значения, либо данная пара встретилась только в одном из документов.

Для того, чтобы выяснить, какая ситуация имеет место, необходимо, предварительно задавшись некоторой доверительной вероятностью, выдвинуть статистическую гипотезу о значении вероятности случайного события.

Если $p_i^* > 0.5$, то выдвигается следующая пара основной и альтернативной статистических гипотез:

$$\begin{aligned} H_0 &: p_i = 1; \\ H_1 &: p_i < 1. \end{aligned} \qquad (2)$$

Если $p_i^* < 0,5$, то выдвигается следующая пара основной и альтернативной статистических гипотез:

$$\begin{aligned} H_0 &: p_i = 0; \\ H_1 &: p_i > 0. \end{aligned} \qquad (3)$$

Схема проверки такого рода статистических гипотез подробно изложена в [7].

Если в (2) основная статистическая гипотеза находит свое подтверждение, то текстовый фрагмент, соответствующий атомарному объекту $\bar{\psi}_i$, будет идентифицирован как атрибут данных с вероятностью $p_i = 1$. Если в (3) основная статистическая гипотеза находит свое подтверждение, то текстовый фрагмент, соответствующий атомарному объекту $\bar{\psi}_i$, будет идентифицирован как атрибут данных с вероятностью $p_i = 0$, т.е. он в этом случае будет идентифицирован как значение атрибута данных.

Если же в (2) или в (3) основная статистическая гипотеза не нашла своего подтверждения, то принимается решение о том, что атомарный объект соответствует текстовому фрагменту документа, являющемуся атрибутом данных с вероятностью $p_i = p_i^*$. И, наконец, если некоторый атомарный объект не фигурирует в отношении ассоциации A, то предварительно принимается решение о том, что он соответствует текстовому фрагменту документа, являющемуся атрибутом данных с вероятностью $p_i = 0,5$ и окончательное решение откладывается до этапа кластеризации анализируемой совокупности текстовых документов.

## Выводы

В данной работе был предложен метод, позволяющий классифицировать обособленные текстовые фрагменты слабоструктурированных текстовых документов как атрибут данных или как значение атрибута данных. Метод основан на сопоставлении абсолютных частот встречаемости текстовых фрагментов, находящихся между собой в отношении ассоциации. Предложен также механизм, основанный на проверке статистических гипотез, позволяющий в некоторых

случаях устранить возникающую неоднозначность в ходе процесса классификации. **Научной новизной работы** является следующее: впервые предложен метод, позволяющий классифицировать обособленные текстовые фрагменты слабоструктурированных текстовых документов как атрибут данных или как значение атрибута данных, основанный на сопоставлении абсолютных частот встречаемости текстовых фрагментов в совокупности документов.

## Список литературы

*1. Губин В.А. Слабоструктурованные текстовые документы как источники данных / В.А. Губин // Бионика интеллекта. – Х.: ХНУРЕ, 2010. – № 3 (74). – С. 109-111.*

*2. Губин В.А. Модель слабоструктурированных текстовых документов / В.А. Губин // Системи управління, навігації та зв'язку. – К.: ЦНДІ НіУ, 2010. – Вип. 4 (16). – С. 213-215.*

*3. Paul Viola. Learning to Extract Information from Semi-structured Text using a Discriminative Context Free Grammar / Paul Viola, Mukund Narasimhand // Proceedings of The 28th Annual International ACM SIGIR Conference Salvador, Brazil, August 15 to 19, 2005.*

*4. Yuk Wah Wong. Scalable Attribute-Value Extraction from Semi-Structured Text / Yuk Wah Wong, Dominic Widdows, Tom Lokovic, Kamal // Proceedings of ICDM Workshop on Large-scale Data Mining: Theory and Applications. December 6-9, Miami, FL,USA, 2009.*

*5. Sujith Ravi. Using structured text for large-scale attribute extraction / Sujith Ravi, Marius Paşca // Proceeding of the 17th ACM conference on Information and knowledge management. ACM New York, NY, USA, 2008.*

*6. Сигорский В.П. Математический аппарат инженера / В.П. Сигорский. – М.: Техника, 1977. – 768 с.*

*7. Кобзарь А.И. Прикладная математическая статистика: для инженеров и научных работников / А.И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.*

*Поступила в редколлегию 22.02.2011*

**Рецензент:** д-р техн. наук, проф. С.Г. Удовенко, Харьковский национальный университет радиоэлектроники, Харьков.

### КЛАСИФІКАЦІЯ ТЕКСТОВИХ ФРАГМЕНТІВ СЛАБОСТРУКТУРОВАНИХ ТЕКСТОВИХ ДОКУМЕНТІВ ЯК АТРИБУТ ДАНИХ АБО ЯК ЗНАЧЕННЯ АТРИБУТУ ДАНИХ

В.О. Губін, Ю.Ю. Шевякова

*Запропонований підхід, що дозволяє віднести той або інший відособлений текстовий фрагмент слабоструктурованого текстового документа до атрибуту даних або до значення атрибуту даних. При цьому передбачається, що аналізується сукупність слабоструктурованих текстових документів одного вигляду. Даний підхід заснований на зіставленні абсолютних частот тієї, що зустрічається текстових фрагментів в сукупності документів.*

***Ключові слова:*** *слабоструктурований документ, дані, атрибут даних, значення даних.*

### THE APPROACH ALLOWS US TO CLASSIFY TEXT FRAGMENTS SEMI-STRUCTURED TEXT DOCUMENTS AS A DATA ATTRIBUTE OR VALUE OF THE DATA ATTRIBUTE

V.A. Gubin, Yu. Yu Shevyakova

*Approach, allowing to deliver one or another isolated text fragment of semi-structured of text document to the attribute of data or to the value of attribute of data, is offered. It is thus assumed that is analysed aggregate of semi-structured of documents of texts of one kind. The examined approach is based on comparison of absolute frequencies of met of fragments of texts in the aggregate of documents.*

***Keywords:*** *semi-structured document, information, attribute of data, value of information.*