Метод извлечения таблиц из неформатированного текста

А. Е. Хмельнов, А. О. Шигаров
Институт динамики систем и теории управления СО РАН, Иркутск, Россия
e-mail: shigarov@icc.ru

The problem of tables extraction is a part analysis of documents. Different approaches to this problem are usually based on certain media and formats. A heuristic method for a plain text table extraction from an unformatted and formatted documents is considered in this paper. This method uses some particular properties of the statistical tables, and it can also be applied to the tables of the similar structure. Additionally, the model of the table structure is proposed, which allows to transform automatically the contents of the extracted tables into relational tables.

Введение

Извлечение таблиц из документов — одна из задач, решаемых в системах анализа и обработки документов. Обзоры работ по извлечению таблиц из документов [1–7], появившиеся за несколько последних лет, показывают растущий интерес к данной проблематике.

Как правило, методы извлечения таблиц из документов ориентируются на определенные среды и форматы представления документов, а также на структуру таблиц, которая обычно определяется стандартами и соглашениями, принятыми в той предметной области, где используются эти таблицы. Одной из наиболее часто используемых сред для представления таблиц является неформатированный текст. Например, большое количество таблиц в статистических сборниках Росстата представлено именно в виде неформатированного текста. Другими примерами сред могут служить документы электронных таблиц (например, Excel), документы текстовых процессоров (например, Word), документы в форматах PDF, HTML или растровые изображения, полученные в результате сканирования бумажных документов.

В данной работе представлены основные концепции разработанного нами эвристического метода извлечения таблиц из неформатированного текста. Метод использует особенности структуры статистических таблиц, публикуемых Росстатом. Эти особенности также в полной мере относятся к статистическим таблицам, представленным в государственных статистических отчетах США (www.fedstats.gov), Евросоюза (Eurostat yearbook 2006–2007) и Японии (Statistical Handbook of Japan 2006). Метод может быть применен к подобным таблицам, представленным как неформатированный текст.

В литературе выделяются следующие основные стадии извлечения таблиц:

— обнаружение таблиц в документах;

— сегментация таблиц на отдельные клетки;

— функциональный анализ — определение роли клеток в таблице;

— структурный анализ — определение зависимостей между клетками;

— интерпретация — преобразование табличной информации к требуемому виду.

В работах [8–10] рассмотрены различные методы обнаружения таблиц в неформатированном тексте. В [11] предложен подход к извлечению таблиц из неформатированного текста, в котором реализованы все указанные стадии извлечения таблиц, но при этом используются слишком сильные предположения о структуре обрабатываемых таблиц. Этот подход ориентирован на особенности таблиц из документации, используемой в строительной промышленности, что не позволяет напрямую применить его к статистическим таблицам.

Предлагаемый нами метод позволяет выполнить все стадии извлечения таблиц, результатом последовательного применения которых являются таблицы реляционной базы данных, содержащие данные, извлеченные из исходных таблиц. В настоящей работе описаны все стадии метода, кроме интерпретации, также предлагается модель для промежуточного представления таблиц, которая может далее интерпретироваться. В разд. 1 рассматриваются особенности статистических таблиц. В разд. 2 предлагается теоретико-множественная модель для представления таких таблиц. В разд. 3 описаны используемые в предлагаемом методе процессы обнаружения таблиц в неформатированном тексте, а также структурного и функционального анализа обнаруженных таблиц.

1. Особенности таблиц

Разнообразие всевозможных форм изображения таблиц очень велико. Часто эти формы определяются стандартами и соглашениями, принятыми в той предметной области, где они используются. В данном разделе рассматриваются особенности статистических таблиц Росстата. Такие таблицы используются для подготовки сборников статистики по сельскому хозяйству, экономике, демографии и др.

На рис. 1 показан пример статистической таблицы. Рассматриваемые таблицы состоят из шапки и тела, кроме того, они могут иметь боковик и перерезы, а также название и единицу измерения. Название таблицы и единица измерения находятся вверху табли-

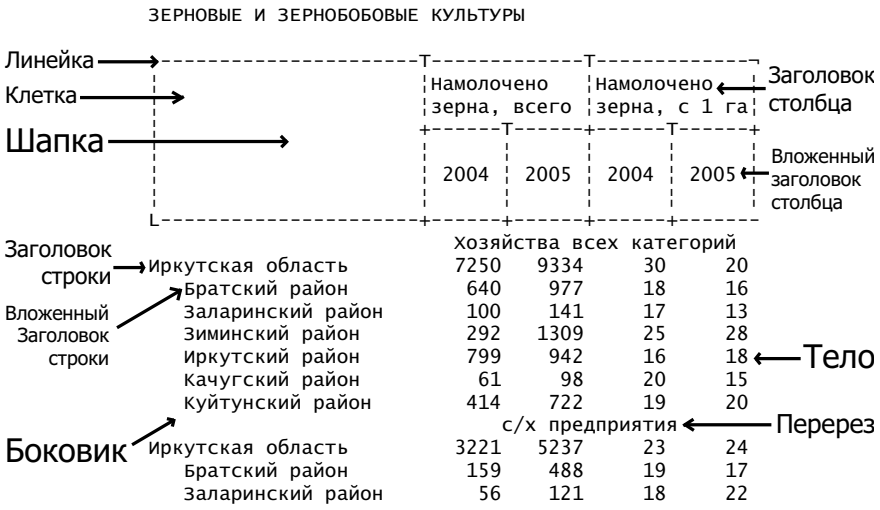

Рис. 1. Пример статистической таблицы

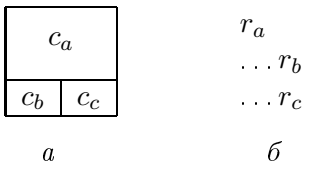

Рис. 2. Примеры вложенности заголовков столбцов (*а*) и заголовков строк (*б*)

цы. Причем название может быть выровнено по левому краю или по центру, а единица измерения — только по правому краю.

После названия и единицы измерения следует шапка таблицы. Она содержит заголовки столбцов, которые обычно выделяются линейками (линиями разграфки), составленными из символов псевдографики или подобных им символов набора ASCII. Пересекаясь, линейки образуют клетки, которые ограничивают отдельные заголовки столбцов. Один или несколько заголовков могут быть вложены в другой заголовок; в этом случае клетки, ограничивающие их, лежат сразу под клеткой, ограничивающей заголовок, в который они вложены. На рис. 2, *а* показан пример вложенности заголовков столбцов: заголовки c_b и c_c вложены в заголовок c_a. Шапка задает ширину для всей таблицы.

Под шапкой располагаются боковик и тело таблицы. Боковик находится слева относительно тела, он состоит из заголовков строк, которые также образуют между собой иерархию вложенности за счет использования отступов. На рис. 2,*б* показан пример вложенности заголовков строк: заголовки r_b и r_c вложены в заголовок r_a. Тело таблицы содержит только числовые значения и специальные символы, указывающие на отсутствие данных.

Внутри тела таблицы могут располагаться перерезы. Они выровнены по центру таблицы и делят ее тело и боковик по горизонтали. При этом заголовки строк, разделенные перерезами, образуют группы. Часто такие группы содержат семантически эквивалентные заголовки строк. К примеру, на рис. 1 видно, что заголовки строк *“Иркутская область”*, *“Братский район”* и *“Заларинский район”* повторяются для перерезов *“Хозяйства всех категорий”* и *“с/х предприятия”*.

В следующем разделе предлагается модель, основанная на данных предположениях об исходной структуре таблицы.

2. Модель таблицы

Особенности компоновки заголовков столбцов позволяют представить шапку в виде дерева, узлами которого являются заголовки столбцов, а ребрами — пары заголовков (c_a, c_b), где c_b — заголовок, вложенный в c_a. Корнем этого дерева является пустой элемент, заголовки самого верхнего уровня — его подузлы. Пусть $C = \{c_1, \ldots, c_n\}$ — множество заголовков столбцов, тогда C^{tree} — дерево заголовков столбцов, представляющее шапку, а $C^{\text{nodes}} = c_0 \cup C$ — множество его узлов, где c_0 — пустой элемент и корень этого дерева.

Подобным образом боковик также можно представить как дерево, в котором заголовки строк являются узлами, а пары заголовков строк, в которых один вложен в другой, — ребрами. Пусть $R = \{r_1, \ldots, r_m\}$ — множество заголовков строк, тогда R^{tree} — множество заголовков строк, представляющее боковик, а $R^{\text{nodes}} = r_0 \cup R$ — множество его узлов, где r_0 — пустой элемент и корень этого дерева.

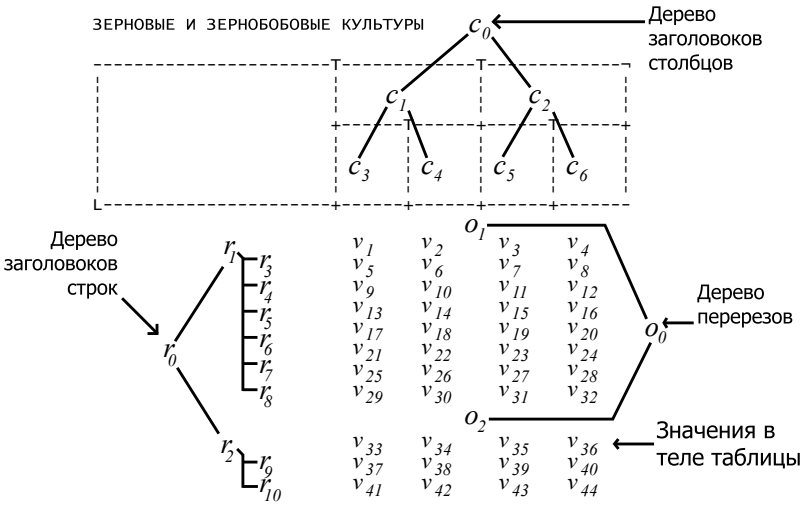

Рис. 3. Пример модели таблицы

Несмотря на то что перерезы не образуют вложенности, их также удобнее рассматривать как одно дерево. Пусть $O = \{o_1, \ldots, o_k\}$ — множество перерезов, тогда O^{tree} — дерево перерезов, а $O^{\text{nodes}} = o_0 \cup O$ — множество его узлов, где o_0 — пустой элемент и корень этого дерева.

На рис. 3 приведен пример модели для таблицы, показанной на рис. 1. Сплошными линиями без стрелок показаны ребра деревьев.

Пусть $V\colon V \subset \mathbb{R}$ — множество значений из тела таблицы, $L \subseteq C^{\text{nodes}} \times R^{\text{nodes}} \times O^{\text{nodes}}$ — подмножество таких элементов из $C^{\text{nodes}} \times R^{\text{nodes}} \times O^{\text{nodes}}$, для которых определено значение $v \in V$. Тогда множество

$$T = \{C^{\text{tree}}, R^{\text{tree}}, O^{\text{tree}}, L \to V\} \tag{1}$$

составляет модель таблицы.

3. Извлечение таблиц

Под извлечением таблицы будем понимать процесс перехода от исходного представления этой таблицы в неформатированном тексте к таблице реляционной базы данных. При этом модель (1) используется как промежуточное представление таблицы, полученное после структурного и функционального анализа обнаруженных в неформатированном тексте таблиц.

3.1. Обнаружение таблиц

Рассматриваемые таблицы являются частью электронных документов. Один документ может содержать несколько таблиц. Между таблицами в документе могут находиться текст и изображения. Обнаружение таблиц в документах — это первая стадия в процессе их извлечения.

Эвристики, используемые при обнаружении таблиц, основаны на ряде предположений о способе представления таблиц в неформатированном тексте. Предполагается, что:

— в тексте каждая таблица имеет последовательно расположенные строки, т. е. не может перемежаться посторонним текстом;

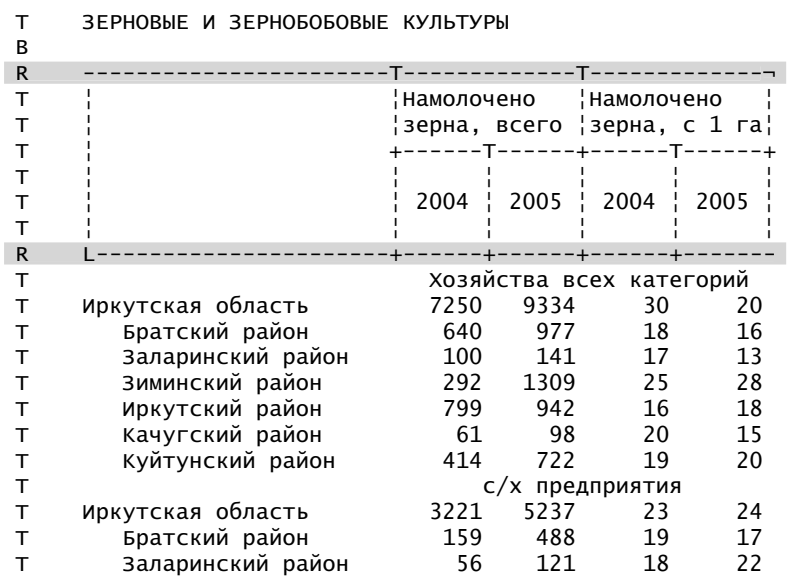

Рис. 4. Фрагмент текста, слева строк указан их тип

— в одной строке могут быть расположены элементы только одной таблицы, в этом случае таблица полностью заполняет строку;

— шапка должна быть обозначена сверху и снизу горизонтальными линейками, каждая из которых занимает отдельную строку и составлена преимущественно из одного символа.

Весь текст документа проходится построчно сверху вниз, начиная с первой строки. Каждая строка текста классифицируется как принадлежащая к одному из трех типов:

B — строка содержит только пробельные символы;

R — более половины строки заполнено одним непробельным символом, т. е. символы в этой строке составляют горизонтальную линейку;

T — строка с текстом, отличная от строк типа B и R.

На рис. 4 показан фрагмент текста, содержащий одну таблицу, слева каждая строка текста отмечена меткой, указывающей ее тип.

В тексте ищутся совокупности строк, типы которых составляют последовательности, вида

$$\{T,\dots,T,B,R,T,\dots,T,R,T,\dots,T\} \text{ или } \{T,\dots,T,R,T,\dots,T,R,T,\dots,T\}.$$

Каждая такая последовательность определяет отдельную таблицу. В результате выполнения этой процедуры определяются последовательности строк текста, составляющие отдельные таблицы. Причем для каждой таблицы определяется, какие именно строки текста составляют ее отдельные части: название с единицей измерения, шапку и тело с боковиком.

3.2. Структурный и функциональный анализ таблиц

Последовательность действий определения модели (1) из исходного представления таблицы в тексте выглядит следующим образом.

1. Анализ шапки таблицы.

1.1. Выделение клеток в шапке.

1.2. Построение дерева шапки по клеткам.
1.3. Определение столбцов таблицы по клеткам.
2. Анализ тела и боковика таблицы.
2.1. Определение строк таблицы.
2.2. Определение значений в теле.
2.3. Определение перерезов и построение дерева перерезов.
2.4. Построение дерева боковика по строкам таблицы.
2.5. Связывание значений в теле с узлами деревьев шапки, боковика и перерезов.

3.2.1. Анализ шапки таблицы

Клетка в шапке представляет собой прямоугольник в текстовых координатах (где точка — это местоположение одного символа, ордината определяется номером строки, абсцисса — номером символа в строке, а начало координат — это самый левый символ в самой верхней строке текста).

Будем считать, что нижняя линия каждой клетки проходит по линии, ограничивающей шапку снизу, при этом все вложенные в нее клетки — это прямоугольники, лежащие полностью внутри нее. Тогда положение каждой клетки однозначно задается шириной w, высотой h и положением ее левого края x, а дерево клеток можно представить в списке, записи которого упорядочены лексикографически по $(x, -h)$, т. е. если $\text{cell}_1 = \{x_1, h_1, w_1\}$ и $\text{cell}_2 = \{x_2, h_2, w_2\}$ — клетки, то $\text{cell}_1 < \text{cell}_2$, тогда и только тогда, когда либо $x_1 < x_2$, либо $x_1 = x_2$ и $-h_1 < -h_2$.

В этом списке вложенные клетки следуют непосредственно за клеткой, в которую они вложены. Чтобы получить клетки, вложенные в i-клетку, достаточно просмотреть список, начиная с позиции $i+1$ до первой клетки, которая имеет высоту, большую или равную высоте i-клетки.

При выделении клеток в первую очередь определяются символы, составляющие горизонтальные и вертикальные линейки. Для этого используется статистический анализ текста шапки.

После того как определены эти символы, алгоритм выделения клеток проходит слева направо самую нижнюю строку шапки в поиске вертикальных разделителей. После обнаружения вертикального разделителя линия разграфки прослеживается снизу вверх до ее окончания. При этом учитывается, что в точках пересечения с горизонтальными линейками вертикальные линейки могут прерываться, поэтому, если встречается символ, отличный от вертикального разделителя, то процесс определения вертикальной линейки не заканчивается, если через этот символ проходит горизонтальная линейка.

Обнаружение вертикальной линейки с высотой h означает, что все ранее найденные клетки с высотой меньше или равной h заканчиваются не далее текущей позиции. Таким образом, после обнаружения вертикальной линейки определяется ширина тех клеток с высотой меньше или равной h, которые до этого не были ограничены справа. Это также означает, что в данной позиции начинается как минимум одна новая клетка с высотой h, ширина которой пока неизвестна. При этом в конец списка клеток добавляется новая запись. После этого вертикальная линейка прослеживается сверху вниз в поиске горизонтальных линеек, уходящих от нее вправо. Обнаружение каждой такой горизонтальной линейки интерпретируется как признак наличия вложенной клетки, при этом в список клеток добавляется соответствующая запись. Это обеспечивает требуемый по-

рядок расположения клеток в списке, при котором вложенные клетки включаются в список после тех клеток, в которые они вложены.

Далее по клеткам из списка определяются столбцы. Если в клетке из списка нет вложенных клеток, то она определяет столбец — ее левый край и ширина совпадают с левым краем и шириной столбца соответственно. Если в клетке $\text{cell}_0 = \{x_0, h_0, w_0\}$ есть вложенные клетки $\text{cell}_1 = \{x_1, h_1, w_1\}$, ..., $\text{cell}_p = \{x_p, h_p, w_p\}$, то она также определяет один столбец, если выполняются следующие условия:

$$w_0 > \sum_{i=1}^{p} w_i;$$

$$x_i + w_i = x_{i+1},\ i = \overline{1, p-1};$$

$$\text{либо } x_0 < x_1,\ x_0 + w_0 = x_p + w_p,\ \text{либо } x_0 = x_1,\ x_0 + w_0 > x_p + w_p.$$

Дерево вложенности клеток определяет дерево заголовков столбцов C^{tree} из модели (1). При этом учитываются только те клетки, внутри которых есть заголовки.

3.2.2. Анализ тела и боковика таблицы

Для того чтобы извлечь тело, боковик и перерезы, алгоритм проходит сверху вниз строки текста между шапкой и концом таблицы. В каждой строке выделяются числа и для каждого числа определяется столбец, которому оно принадлежит. При этом допускается, чтобы диапазон x-координат числа частично выходил за диапазон x-координат столбца, не превышая заранее определенной пороговой величины. С другой стороны, попадание нескольких чисел из одной строки в один столбец или одного числа в несколько столбцов интерпретируется алгоритмом как полное несоответствие всей строки структуре шапки. Из строк, соответствующих структуре шапки, выделяются числа, образующие множество значений тела таблицы — V из модели (1).

Самый левый столбец интерпретируется как боковик, т. е. числа, которые туда попадают, также рассматриваются как части заголовков строк. В ходе обнаружения значений тела также определяются заголовки строк исходя из того, что диапазон x-координат заголовка строки не должен выходить за пределы диапазона x-координат боковика.

Возможна ситуация, когда заголовок строки таблицы не помещается в одну строку текста. В этом случае предполагаем, что значения строки таблицы, относящейся к данному заголовку, всегда будут размещаться в последней строке текста, составляющего данную строку таблицы. Поэтому строки текста, в которых полностью отсутствуют значения, объединяются со следующими за ними строками текста, в которых значения присутствуют, при условии, что у этих строк совпадает отступ (так несколько строк текста образуют одну строку таблицы).

Для того чтобы извлечь перерезы, проходя по строкам текста, ищут последовательности идущих подряд строк текста, которые выровнены по центру таблицы. Если такая последовательность будет найдена, то весь ее текст составляет один перерез. Все найденные перерезы составляют дерево перерезов O^{tree} из модели (1).

По отступам от левого края таблицы определяются вложенность обнаруженных заголовков строк и дерево R^{tree} из модели (1). При этом предполагается, что заголовки строк одного уровня имеют одинаковый отступ от левого края таблицы и на каждом более низком уровне отступ увеличивается. Строки текста, занимаемые вложенным заголовком, всегда расположены ниже строк текста, занимаемых заголовком, в который он вложен.

Каждое найденное число в теле таблицы связано со строкой, столбцом и перерезом таблицы. По этому расположению чисел внутри таблицы для каждого значения $v \in V$ определяется пара $\big((c, r, o), v\big) \in C^{\text{nodes}} \times R^{\text{nodes}} \times O^{\text{nodes}} \to V$, где c — заголовок столбца, r — заголовок строки, o — перерез, с которыми связано значение v. Таким образом, строится множество $L \subseteq C^{\text{nodes}} \times R^{\text{nodes}} \times O^{\text{nodes}}$. Приходим к модели (1).

Таблица, представленная как модель (1), может быть подвергнута дальнейшим преобразованиям для приведения к отношению в реляционной модели.

Заключение

При решении многих научных и практических задач анализа и обработки данных требуется наполнять базы данных данными из таблиц, содержащихся в электронных документах. При этом таблицы в документах могут иметь различные сложные структуры. Примером таких таблиц могут служить статистические таблицы, рассматриваемые в данной работе. Их нельзя преобразовать в таблицы реляционной базы данных стандартными программными средствами. Методы извлечения таблиц из документов (в частности, из неформатированного текста) позволяют ускорить процесс наполнения баз данных для некоторых задач, сделать его полуавтоматическим.

Статистические таблицы являются классом таблиц, имеющих схожие структуры и типы заголовков. Это позволило сделать некоторые предположения об этих таблицах и сформулировать эвристики, используемые предлагаемым в данной работе методом извлечения таблиц.

На основе предлагаемого метода авторами разработано программное обеспечение для автоматизации наполнения баз данных. С его помощью выполнены работы по наполнению статистической базы данных по сельскому хозяйству Иркутской области. Было обработано около 2800 таблиц, представленных как неформатированный текст в документах Microsoft Word и ASCII-текстах, из которых извлечено более 21 000 показателей и более 300 000 значений.

Авторам неизвестно об использовании подобных методов автоматизации наполнения статистических баз данных на территории России и о существовании распространяемого коммерческого или свободного программного обеспечения, решающего подобные задачи.

Список литературы

[1] Handley J.C. Document recognition // Electronic Imaging Technology, chapter 8. IS&T/SPIE Optical Eng. Press, 1999.

[2] Embley D.W., Hurst M., Lopresti D., Nagy G. Table-processing paradigms: a research survey // Intern. J. on Document Analysis and Recognition. 2006. Vol. 8, N 2. P. 66–86.

[3] Embley D.W., Lopresti D., Nagy G. Notes on contemporary table recognition // Proc. 7th Intern. Workshop on Document Analysis Systems (DAS). Springer, 2006. P. 164–175.

[4] Hurst M. The Interpretation of Tables in Texts: Ph. D. Thesis. N.Y.; Berlin: Univ. of Edinburgh, 2000.

[5] Lopresti D., Nagy G. A tabular survey of automated table processing // Lecture Notes in Computer Sci. 2000. Vol. 1941. P. 93–120.

[6] Lopresti D., Nagy G. Automated table processing: An (opinionated) survey // Third IAPR Intern. Workshop on Graphics Recognition. Jaipur, India, 1999. P. 109–134.

[7] Zanibbi R., Blostein D., Cordy J.R. A survey of table recognition: Models, observations, transformations, and inferences // Intern. J. on Document Analysis and Recognition. 2004. Vol. 7, N 1. P. 1–16.

[8] Hu J., Kashi R., Lopresti D., Wilfong G. Medium-independent table detection // Document Recognition and Retrieval VII. IS&T/SPIE Electronic Imaging, San Jose, 2000. P. 291–302.

[9] Tupaj S., Shi Z., Chang C.H., Alam H. Extracting Tabular Information From Text Files. 1996. `http://citeseer.nj.nec.com`

[10] Pinto D., McCallum A., Wei X., Croft B. Table extraction using, conditional random fields // 26th Annual Intern. ACM SIGIR, Conf. on Research and Development in Information Retrieval, 2003.

[11] Douglas S., Hurst M., Quinn H. Using natural language processing for identifying and interpreting tables in plain text // 4th Annual Symp. on Document Analysis and Information Retrieval. Las Vegas, 1995. P. 535–546.

Поступила в редакцию 25 января 2008 г.