КОМПЬЮТЕРНЫЕ НАУКИ

УДК 519.713

МОДЕЛИ ПРЕДСТАВЛЕНИЯ ДАННЫХ В WORLD WIDE WEB

БОНДАРЕНКО М.Ф., КОРЯК А.С., РОШКА С.А., ТОМИЧ И.В.

Предлагается классовая модель данных. Представление текстовых данных коллекции Web-страниц с помощью этой модели позволяет применять методы исследования взаимного расположения значимых элементов (слов, фраз, термов) коллекции и алгоритмы кластерного анализа для понятийно-ориентированных методов технологии text mining.

1. Введение

В последнее время наблюдается присутствие огромного количества HTML документов в сети – страниц сайтов. Пользователь нуждается в анализе коллекции документов в целях определения тематических направлений, принадлежности их к определенным тематикам, связей между ними. С другой стороны, пользователю может потребоваться найти специальную часть информационного содержания.

Цель text mining (интеллектуального анализа данных) – автоматизировать насколько возможно преобразование найденной информации (линейной) в знания или обобщенные шаблоны (скрытую информацию). Эта новая преобразованная информация сможет применяться непосредственно пользователем для обзора интересуемой области и получения релевантных решений. Одним из основных применений является мониторинг ситуации в WWW (World Wide Web) по интересуемым тематическим направлениям [1].

Text mining включает многие шаги из области обнаружения знаний, от очистки данных до визуализации знаний. Доминирующими направлениями исследований в text mining являются анализ текста, текстовая интерпретация, документная категоризация и документная визуализация. Для извлечения понятийно-ориентированных знаний из коллекций документов HTML (Web-страниц) необходимо привлекать технологии text mining и KDT (Knowledge Discovery from Text – извлечение знаний из текста).

Процесс реализации технологий text mining или KDT можно разделить на три этапа: выбор данных и предобработка, размещение в хранилище данных; анализ данных - извлечение неизвестной информации и преобразование линейных данных в шаблоны; представление данных (визуализация) пользователю.

Методы, используемые в анализе данных [2], можно классифицировать на 3 группы: классификация - отображение данных в определенные классы или построение кластеров соответственно подобию данных; определение (взвешивание) зависимостей между переменными, отношений между полями, временных зависимостей или последовательностей, регрессий; преобразование - реферирование.

Теговую структуру HTML документа в задачах анализа текстовых данных можно использовать различными способами. Извлекаемые шаблоны теговых данных применяются в технологиях text mining и KDT, что способствует представлению текста, заключенного в такие шаблоны, моделям данных. Построение моделей данных в текстах – Data Warehousing [3] - актуальное направление исследований, позволяющее структурировать текстовую информацию в целях дальнейшего извлечения и представления знаний из текста, применять технологии информационного поиска в документах, извлекать фактографическую информацию.

Многие предыдущие методы для автоматической обработки Web-ориентированных данных концентрируются на их структурных характеристиках. Они главным образом основаны на специальных грамматиках для получения четкой основной структуры или используют схемы, где представляют HTML страницу как объект с атрибутами, такими как URL, title, author.

Для многозначного и правильного применения Web-ориентированных данных требуется использование их организационной и смысловой информации. Эта информация, которая называется контекстной, предоставляет основу для определения отношений между данными и аспектов их описания. Имеются отличия между структурными и семантическими метаданными. Структурные метаданные представлены информацией, которая описывает организацию и структуру отмеченных данных, т. е. информацию о формате, используемом типе данных, синтаксических отношениях между ними. Семантические данные предоставляют информацию о значениях доступных данных и их отношениях.

Целью исследования является представление новой информационной модели данных, применяемой на этапе предварительной обработки документов, с использованием теговой структуры HTML документа. Новый подход к применению структуры документов в информационных моделях данных позволяет разбить текст на семантически значимые последовательности, что определяет базовые классы контекстов употребления ключевых сущностей. Модель данных определяет возможности дальнейшей обработки извлекаемой информации, формирует пространство признаков, которые будут инициирующими для применения технологий text mining и KDT.

Цель данной работы – определение тематических направлений в коллекции нечетко структурированных данных (WWW-страницах, извлекаемых по запросу к поисковым машинам).

Для достижения поставленных целей необходимо решить следующие *задачи*: определить базовые классы из набора текстовых данных, способствую-

щие эффективному применению методов технологии text mining и KDT; произвести натуральную языковую обработку текста для выявления дополнительных характеристик на множестве классов; применить методы кластерного анализа данных.

2. Обзор методов обработки данных в WWW

Основным препятствием для использования внутренней структуры Web-документов является отсутствие схемы или типов данных и возможных ошибок, которые появляются по этой причине. Задача осуществления запросов к данным, чья структура неизвестна или иррегулярная, относится к языкам запросов для нечетко структурированных данных, Lorel [4] и UnQl [5]. Под нечеткой структурой подразумевается, что данные могут иметь некоторую структуру, не жесткую, не регулярную и не полную, как в традиционных системах управления базами данных [6].

Lorel – Lightweight Object Repository – это проект Stanford для предоставления удобного и эффективного хранилища, осуществляющего запросы и обновление нечетко структурированных данных [4]. Все данные в Lorel базе распределяются соответственно Object Exchange Model (OEM) [7]. В OEM каждый объект является атомарным (integer, string, image) или комплексным. Комплексный объект состоит из коллекции отмеченных подобъектов. Поэтому OEM объект может быть выражен через корневой граф, с объектами-вершинами и метками-дугами, представляющими отношения объект-подобъект. Запросы формируются основываясь на в*ыражении пути*, который является последовательностью меток, идущих из корня. Lorel поддерживает произвольное распределение атомарных данных в рамках OEM объекта, включая целые, вещественные, строки, графику, аудио, видео, HTML текст, Java апплеты.

UnQl – закрытый язык в отличие от Lorel, был разработан для работы с нечетко структурированными данными AT&T. UnQl основывается на модели, состоящей из корневого, отмеченного графа подобно OEM.

Если WWW рассматривать как большую, графоподобную базу данных, то такая структура применяется для информационного поиска современными поисковыми машинами, где объединяются внутренняя структура Web-страниц и внешняя структура ссылок. На сегодняшний день предлагаются WebSQL [8], WebLog [9], W3QS [10] как языки запросов к Web-данным.

W3QS – (WWW Query System) является проектом для разработки гибкого, декларативного и SQL-подобного языка Web-запросов. Язык поддерживает эффективный и гибкий обработчик запросов, который работает со структурой и контентом WWW узлов и их различными видами данных.

WebSQL – система специализируется на HTML документах и гиперссылках, исходящих из них. Для реализации контентных и структурных запросов предлагается модель WWW как реляционная база данных, сформированная из двух виртуальных отношений:

Document [url, title, text, type, length, modif],

Anchor [base, label, href],

где url является ключевым, а все остальные атрибуты могут быть нулевыми.

WebLog – декларативный язык для Web-запросов, основан на SchemaLog [11]. Концептуальная модель WebLog применяется к HTML информации. Модель для WebLog основана на представлении, что каждый Web-документ содержит гетерогенную смесь информации о приведенной выше теме в тегах описания документа. На практике типичный Web-документ состоит из групп связанной информации, которая пространственно ограничена одной страницей. Информация в пределах каждой такой группы будет гомогенной (однородной). Для примера, информация, заключенная в тег <HR> в документе, будет формировать группу однородной информации. Каждая группа называется rel-infont. Страница является набором rel-infont. Состав rel-infont очень субъективный и их выбор зависит от пользователя. Также rel-infont может иметь атрибуты, обычно это теговые атрибуты и каждый rel-infont имеет свой уникальный идентификатор.

WebOQL – система [12], поддерживающая основные классы операций реструктуризации данных в среде WWW. Основной структурой данных, предоставляемой WebOQL, является гипердерево. Оно упорядочено деревом с отмеченными дугами и двумя типами дуг - внутренними и внешними. Внутренние дуги используются для представления структурных объектов, а внешние – для представления отношений (гиперссылок) между объектами. Дуги помечаются записями. Это дерево может быть построено из HTML файла с помощью специального для HTML программного посредника (wrapper).

ARANEOUS – эта система [13] предлагается для управления и реструктуризации извлекаемых данных из WWW. Atzeni, Mecca, Merialdo представляют специализированную модель данных, названную Araneous Data Model (ADM) и использующую обобщенные структуры, представляемые в Web-сайтах. Модель позволяет один раз описать схему гипертекста WWW в реляционном виде. ADM является странично-ориентированной моделью, в смысле основного построения модели по схеме Web-страницы. Она использует описание структуры набора гомогенных страниц сайта: HTML страница рассматривается как объект с идентификатором, URL и несколькими атрибутами, по одному для каждой релевантной части информации на странице. Схема Web-сайта может рассматриваться как коллекция схем страниц, связанная ссылками. Атрибуты могут быть простыми и комплексными.

STRUDEL – система управления Web-сайтами. Разработана в AT&T и применяет понятия, близкие к системам управления базами данных для процесса построения Web-сайтов. STRUDEL имеет графовую модель и подобна OEM [9]. Графы содержат объекты, или отмеченные вершины, связывающиеся с отмеченными гранями именами атрибутов. STRUDEL также предоставляет коллекции, которые называются наборами объектов, и поддерживает некоторые атомарные типы, которые повсеме-

стно присутствуют в Web-страницах, такие как URL, PostScript, text, image, HTML files.

SIM – менеджер структурированной информации, (Structured Information Manager) [8] является системой содержащих базы данных документов, разработанной для управления многогигабайтной коллекцией документов, содержащей неструктурированный текст (ASCII), структурированный текст (включая SGML и MARC), бинарные объекты (такие как рисунки или видео) и другие виды данных. Центром SIM системы является мощная база данных, которая понимает структуру и контент документов.

WHOM – объектная модель (Warehouse Object Model) [15]. WHOM Web-кортеж представляется множеством направленных графов, каждый из которых состоит из множества вершин и ссылок и соответствует Web-схеме. Модель представления содержимого и структуры Web-документа представляется в иерархично-графовом виде, Web-документ отображается в экземпляр дерева структурных атрибутов. Применяется игнорирование "шумных" элементов в документах. К "шумным" относятся элементы структуры HTML или XML документов, которые не включаются в модель и не несут значимой информации для хранилища данных.

3. Классовая модель представления данных HTML документа

Перед применением методов KDT к документам необходимо заключить данные в определенную структуру [16], свойства которой позволят применять требуемые методы анализа.

Способ использования совместно текстовых данных, размещаемых в документах, и структурной разметки HTML (тегов) определяется моделью данных. Для формирования тематик коллекции документов (смысловых категорий) требуется определить контекст, в котором употребляется устойчивая последовательность (словосочетание). Мы используем теги HTML документов для разделения текстовой части документа на классы. Совокупность классов составит понятие в документе. Понятия помогают ввести набор базовых смысловых категорий в тексте. Нам важно отметить границы определения контекста, в котором будут употребляться устойчивые последовательности. Границами контекста может служить текст всего документа, или определенное расстояние между словами. В нашем случае границы контекста различны, они будут зависеть от схемы использования тегов в документе.

В представленной модели данных мы используем теговую структуру документа для определения информационного контекста, заключенного в извлекаемых шаблонах. В технологиях Web content mining теговая структура HTML документа используется для применения гибких технологий в анализе текстовых данных.

В современных методах обработки текста контекст употребления слов, термов или ключевых выражений определяется их совместным расположением. Границы совместного расположения определяются следующими способами:

– парным совместным расположением термов;

– размером окна – интервалом [-N, N] термов от исследуемого в контексте всего документа;

– с помощью тегов HTML, которые выражают семантический смысл данных, заключенных между ними – это такие теги как meta, author, title, link, метки ссылок.

Мы применяем разделение тегов на классы (группируем), на основании которых будут определяться контекстные шаблоны данных в документах HTML. В таблице представлены теги, включенные в модель. В модели мы полагаем, что контекст в Web-страницах выделяется с помощью тегов форматирования текста [17], исключая теговые атрибуты и их значения, которые относятся к стилевому оформлению документа. С помощью правил распределения данных документа по классам модель отфильтровывает шумные данные.

Теги, используемые в классовой модели данных

№ класса	Теги
001	<p>, , , <wbr>
002	abbr, b, hx, big, blockquote, center, cite, dfn, em, I, q, pre, small, span, strong, tt, u, acronym, font
003	<a>метка
004	Списки – <dd>, <dl>, <dt>, , ,
005	Таблицы - <td>

В нашей классовой модели данных мы используем теги, которые направлены на представление текстовых данных HTML-документа без приемов стилевого оформления Web-страницы, т. е. это теги форматирования, которые используются, чтобы выделить часть текста, и теги для контекстного разделения представляемой информации – это абзац, таблица, списки, метки ссылок, надписи рисунков. Такое использование тегов позволяет от плоского текста перейти к взвешенной системе подачи информации в пределах контекста. Остальные теги удаляются из документа и считаются шумными.

В классы включается текст, который располагается между этими тегами. Есть теги, у которых отсутствует закрывающий тег или он упускается, например
, тогда завершающий ему тег ставится в позиции, где встречается любой открывающий тег такого же класса.

Целью представления в классовой модели является вывод списка понятий для каждого документа, получаемых из схемы размещения классов в документе и текстовых данных, соответствующих каждому понятию в документе.

Мы применяем схему кодирования документов для дальнейшего использования данных из нашей модели на машинно-понятном языке.

Модель представления данных состоит из нескольких этапов. Правила формирования совокупностей классов: выделяем текст, который расположен между открывающим и закрывающим тегом одного из

перечисленных классов. Рассматриваем теги от текста, заключенного между ними. Если открывающие теги идут подряд в одном классе, то им в соответствие ставится номер надлежащего класса. Если далее идет тег, соответствующий номеру другого класса, то он включается в модель, и т. д., пока последовательность тегов не прервется тегом, не включенным в модель.

Извлекаем содержимое классов в следующем формате:

1w №класса **K** Knum **Id** Idnim **&var** &varnum **E** [[]] **0w** №класса,

где **1w** – признак открывающего тега; №класса – номер класса, к которому относится тег в нашей модели данных; **K** – признак наличия текста; Knum – количество термов в тексте, заключенном между этими тегами; **Id** – признак идентификатора текстовой последовательности; Idnim – идентификатор последовательности текста в документе (номер извлеченной последовательности в порядке обнаружения в документе); **&var** – признак ссылки на текст в документе; &varnum – ссылка на текст в документе, которая соответствует номеру уникального ключа записи в базе данных; **E** – признак завершения описания текстовой последовательности; [[]] – возможно вложение классов; **0w** – признак закрывающего тега.

Часть текста, ограниченная тегами модели, представляется в формате:

K Knum **Id** Idnim **&var** &varnum **E.**

Пример следующего фрагмента HTML документа:

<P class=par>Афера и мошенничество — по сути одно и то же. Статья 159 Уголовного кодекса Российской Федерации трактует это как “хищение чужого имущества или приобретение права на чужое имущество путем обмана или злоупотребления доверием”;

<P class=par>И это самое “хищение чужого имущества путем обмана или злоупотребления доверием” пышно цветет на ниве всеобщей любви к “халяве” и неистребимой веры в чудеса;

<P class=par>Чтивший

<Ahref=”http://www.aferizm.ru/YK/YK_RF.htm”> Уголовный кодекс</A> О.И. Бендер знал четыреста относительно честных способов отъема имущества у сограждан. Но прогресс в индустрии жульничества и обмана не стоит на месте — современные жулики совершенствуют старые и изобретают новые приемы надувательства. И сколько этих приемов, относительно честных и абсолютно бесчестных, не знает даже Главное управление по борьбе с экономическими преступлениями МВД России.</P>

<P class=par><IMG height=100 alt=”a_raduga.gif (7455 bytes)” hspace=1 src=”Аферы.files/a_raduga.gif”

width=125 align=right vspace=1 border=0>И хотя у народа заметно убавилось доверчивости к “пирамидам”, наперсточникам, “ <A href=”http://www.aferizm.ru/moshen/lohotron.htm”>лохотронам</A>”, и уже не каждая помойка кажется “Полем Чудес” (не путать с телепередачей), новые технологии жульничества и вариации известной байки сменивших имидж кота Базилио и лисы Алисы находят горячий отклик в умах и, что самое главное для аферистов, карманах сограждан.</P>

документ разобьется на такие понятия:

1w1K32id1&var1E0w1

1w1K24id2&var2E0w1

1w1K56id3&var3E1w3K2id4&var4E0w30w1

1w1K50id5&var5E1w3K1id6&var6E0w30w1

и также текст закодируется такими последовательностями, соответственно формату:

K32id1&var1E

K24id2&var2E

K56id3&var3E, K2id4&var4E

K50id5&var5E, K1id6&var6E.

Далее полученные списки совокупностей классов преобразовываются к следующему формату:

cl№класса (Tseq, [], ...),

где Tseq – последовательность термов, записанная в формате, представленном ранее для текста; [], ... – возможно вложение классов.

Мы получаем совокупности классов, которые определяем как **понятия**. Понятия в нашей модели позволяют определить базовые совокупности контекста употребления слов в коллекции документов. В модели сохраняется порядок следования совокупностей классов.

Из фрагмента HTML документа мы получаем следующие классовые совокупности:

cl1(K32id1&var1E);

cl1(K24id2&var2E);

cl1(K56id3&var3E, cl3(K2id4&var4E));

cl1(K50id5&var5E, cl3(K1id6&var6E)).

Классовая модель обладает следующими свойствами, которые будут применяться в дальнейшем в наших исследованиях к задаче извлечения знаний из текста.

Свойство 1. Текст разбивается на близкие по смыслу текстовые совокупности, дающие интервал для дальнейшего применения в исследовании совместного расположения ключевых последовательностей (термы, ключевые фразы).

Свойство 2. Получаемые классовые совокупности представляют разбиения на смысловые категории, образуемые из базовых смысловых категорий (их пять - классы в модели), которые можно применять для получения характеристической оценки понятия (меры веса) в пространстве признаков.

Характеристика понятия выражается матрицей. Пусть ClN – количество базовых классов в модели, тогда $\|Track_{i,j}\|$ i,j=1..ClN является характеристической матрицей понятия P, образуемого множеством классов модели Cl. $Track_{i,j} = 1$, если класс j включен в класс i понятия, 0 - в противном случае.

Свойство 3. Выявление совпадающих или подобных фрагментов текста, по признаку &var. Определение подобных фрагментов текста мы дадим в следующей статье, которая будет описывать метод извлечения устойчивых словосочетаний в коллекции документов и свойства, полученные по классовой модели фрагментов текста.

Свойство 4. Понятия заключают в себе наборы слов текста – последовательности, которые мы можем перенумеровать по отношению ко всей коллекции документов. Понятия мы также можем перенумеровать по всей коллекции документов. Пусть Seq – множество последовательностей всей коллекции документов, s – количество последовательностей, P – множество понятий коллекции документов, p – количество понятий в коллекции, Track – характеристическая матрица каждого понятия p в коллекции, CS – вектор $CS=(P_1, .., P_p)$, $P_i=1$, если Seq_s встречалась в понятии P_i, i=1..p.

Характеристиками модели для каждой Seq_s – последовательности с номером s является набор множеств: {Seq(s), CS(s)}, а для каждого понятия P_p – понятие с номером P – {Track(p), CS(p)}.

4. Выводы

В статье *проведен обзор* существующих методов использования тегов HTML документов и обзор моделей представления данных в WWW, которые на сегодняшний день применяются в технологиях text mining и KDT.

Научная новизна отражена в предлагаемой модели данных для обработки коллекции Web-страниц, которая использует теги HTML разметки документов и помогает употребить эффективный механизм для определения контекста применения ключевых выражений текста.

Практическая значимость определяется возможностью дальнейшего применения модели при решении задач извлечения понятийно-ориентированных знаний из текста – определения тематического состава коллекции Web-документов с помощью методов кластерного анализа.

В сравнении с аналогами получили, что предыдущие методы для автоматической обработки Web-ориентированных данных концентрируются на их структурных характеристиках. Они главным образом основаны на специальных грамматиках для получения четкой основной структуры или используют схемы, где представляют HTML страницу как объект с атрибутами, такими как URL, title, author. Такая организация данных показывает хорошие результаты применения в задачах информационного поиска. Предлагаемая модель использует тегову структуру документов в целях формирования свойств и их значений для интеллектуального анализа данных, группирует данные в контексте их смыслового употребления.

Литература: 1. *Mothe J.* Internet-Based Information Discovery: Application to Monitoring of Science and Technology, Research in Official Statistics. 1998. № 1. P.17-30. **2.** *Fayyad, U.M., Piatetsky-Shapiro G., Smith G., P.* The KDD process for extracting useful knowledge from volumes of data, In Communications of the ACM. 1996. Vol.39, № 11. P. 27-34. **3.** *Sourav S. Bhowmick, Sanjay Madria, Wee Keong Ng.*: Representation of Web Data in a Web Warehouse, The Computer Journal. 2003.Vol. 46, № 3. P. 30-62. **4.** *Abiteboul, S., Quass, D., McHugh, J., Widom, J. and Weiner, J.* The Lorel query language for semistructured data, Int. J. of Digital Libraries. 1997. № 1. P. 68-88. **5.** *Buneman, P., Davidson, S., Hillebrand, G. and Suciu, D.* A query language and optimization techniques for unstructured data, In Proc. ACM SIGMOD Int. Conf. on Management of Data, Canada, June, ACM Press, New-York. 1996. P. 505-516. **6.** *Buneman, P.* Structured data, In Proc. Int. Conf. on Principles of Database Systems, Tucson, Arizona. 1997. May 12-14, ACM Press, New-York. P.117-121. **7.** *Papakonstantinou Y., Garcia-Molina, H. and Widom J.* Object exchange across heterogeneous information sources, In Proc. ICDE 95, Taipei, Taiwan, March 6-10, IEEE Computer Society, Los Alamitos, CA. 1995.P. 251-260. **8.** *Mendelzon A.O., Mihaila G.A. and Milo T.* Querying the World Wide Web, In Proc. Int. Conf. On Parallel and Distributed Information Systems (PDIS'96), Miami, FL, December 18-20, IEEE Computer Society, Los Alamitos, CA.1996. P.80-91. **9.** *Lakshman L.V.S., Sadri F. and Subramanian I.N.* A declarative language for querying and restructuring the Web, In Proc. Sixth Int. Workshop on Research Issues in Data Engineering, New Orleans, LA, February 26-27, IEEE Computer Society, Los Alamitos, CA. 1996. P. 12-21. **10.** *Konopnicki D. and Shmueli O.* Information gathering in the World-Wide Web: the W3QL query language and W3QS system, Theory of database Systems (TODS). 23. 1998. P. 369-410. **11.** *Lakshman L.V.S., Sadri F. and Subramanian I.N.* SchemaSQL: a language for interoperability in relational multi-database systems, In Proc. 22nd Int. Conf. on Very Large Data Bases, San Fransisco, CA.1996. September 3-6. P. 239-250. **12.** *Arocena G. and Mendelzon A.* WebOQL: restructuring documents, databases and Webs, In Proc. ICDE 98, Orlando, FL, February 23-27, 1998. P. 24-33. **13.** *Atzeni P., Mecca G. and Merialdo P.* To weave the Web, In Proc. 22nd Int. Conf. on Very Large Data Bases, Athens, Greece.1997. August 25-29, Morgan Kaufmann, San Fransisco CA. P. 239-250. **14.** *Sack-Davis R. and Kent A. J.* The Structured Information Manager (SIM), In Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Melbourne, Australia, August 24-28, ACM Press, New York. 1998.P. 387. **15.** *Bhowmick S.S., Ng W.K. and Madria S.K.* Schemas for web data: a reverse engineering approach, Data and Knowledge Eng. J. (DKE). 2001.Vol. 39. P. 105-142. **16.** *Feldman and Dagan, I.*: Knowledge Discovery in Textual Databases (KDT), In Proceedings of the 1st International Conference on Knowledge Discovery (KDD-95), Montreal, Aug 1995. P. 112-117. **17.** http://www.w3c.org

Поступила в редколлегию 02.11.2004

Рецензент: д-р техн. наук, проф. Путятин В.П.

Бондаренко Михаил Федорович, д-р техн. наук, профессор, ректор ХНУРЭ. Научные интересы: разработка теоретических основ создания и использования систем искусственного интеллекта различного назначения; компьютерная лингвистика и лексикографические системы; интеллектуальный анализ текста. Адрес: Украина, 61166, Харьков, пр. Ленина, 14.

Коряк Алексей Сергеевич, аспирант ХНУРЭ, каф. ПОЭВМ. Научные интересы: разработка методов интеллектуального анализа в сети World Wide Web; интеллектуальный анализ текста. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, e-mail: Alex@smit.com.ua

Рошка Светлана Александровна, аспирантка ХНУРЭ, каф. ПОЭВМ. Научные интересы: разработка методов интеллектуального анализа в сети World Wide Web; интеллектуальный анализ текста. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, e-mail: wstudio@ua.fm, ICQ 304667027.

Томич Игорь Владимирович, студент факультета КИУ ХНУРЭ. Научные интересы: разработка методов интеллектуального анализа в сети World Wide Web; интеллектуальный анализ текста. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, e-mail: makoomazan@mail.ru