УДК 004.853

О.М. Почанский
ХНУРЭ г. Харьков, Украина, pochansky.oleg@yandex.ru

ИЗВЛЕЧЕНИЕ ЧАСТИЧНО-СТРУКТУРИРОВАННОЙ (ЗНАЧИМОЙ) ИНФОРМАЦИИ ИЗ ДИНАМИЧЕСКИХ WEB-ДОКУМЕНТОВ

В данной статье рассматривается модель системы, которая отвечает за извлечение значимой информации из web-документов. В основе модели лежит метод разделения web-страницы на структурные блоки. Каждый из них должен проходить проверку на наличие шумов. Те данные, которые прошли проверку и не содержат шумов, сохраняются в базе знаний системы.

ИЗВЛЕЧЕНИЕ ЗНАНИЙ, БАЗА ЗНАНИЙ, ЗНАЧИМАЯ ИНФОРМАЦИЯ, ШУМ, ОНТОЛОГИЯ, ДИНАМИЧЕСКИЙ WEB-ДОКУМЕНТ

Введение

Проблеме извлечения значимой информации из web-документов посвящается огромное количество научных статей и прикладных разработок в течение длительного промежутка времени (начиная с конца 20 века и по наши дни). Но в большинстве существующих решений присутствует ряд ограничений или допущений, связанных с написанием эффективных алгоритмов работы специализированных систем, способных проанализировать и обработать почти любой источник, содержащий необходимые данные [1]. В первую очередь, это связано со сложностями понимания ими сути значимой информации и отличием ее от другой информации, которая не удовлетворяет требованиям или критериям, поставленным перед системой. Таким образом, это позволит утверждать о том, что данная проблема в полной мере не решена, а значит, исследование в данной области не утратило своей актуальности.

1. Причины возникновения данной проблемной области

Попробуем разобраться в причинах возникновения указанной проблемы. Итак, для начала остановимся на определении общего понятия термина информация. Данный термин означает сведения, передаваемые источником получателю (приемнику). Он всегда связан с материальным носителем, с материальными процессами и имеет некоторое представление. Информация, представленная в какой-либо форме, называется сообщением. Сообщения представляются в виде сигналов и данных. Сигналы используются для передачи информации в пространстве между источником и получателем, а данные — для хранения (т. е. для передачи во времени) [2].

В нашем случае значимая информация будет рассматриваться в виде произвольного набора данных, сгруппированных в рамках заданной общей предметной области. По способу упорядочивания данных в различных типах документов их можно разделить на три вида: неструктурированные, частично-структурированные и структурированные [3].

Как правило, электронные документы составлены в произвольной форме на естественном языке и содержат неструктурированные данные. Примером такого документа может быть статья с произвольной тематикой, опубликованная в журнале или газете. Частично-структурированные данные содержат в основном электронные Web-документы различных расширений (php, aspx, jsp, htm и т.д.), оформленные в текстовом формате HTML, где они описываются с помощью специальных тэгов. И наконец, структурированные данные содержат XML-документы и базы данных [1]. Их относят к данному типу благодаря наличию специфических инструментов. В случае XML это – DTD (преамбула документа, где определяются его компоненты и структура) и XML schema (язык описания структуры XML документа) [1]. В случае базы данных это – функциональная особенность программной оболочки, в которой она была создана, и специализированный язык запросов – SQL, с помощью которого можно обратиться к любому существующему элементу базы данных.

Следовательно, при исследовании вопроса извлечения значимой информации необходимо учитывать не только соответствие извлекаемых данных одной предметной области, но и тип источника, в котором они находятся.

При этом к основным критериям значимости любой информации можно отнести: актуальность, достоверность, полноту и «чистоту» находящихся в ней данных. Первые три критерия можно отнести к субъективным понятиям, которые могут быть выявлены экспериментальным путем. В свою очередь, последний критерий можно отнести к объективным характеристикам, так как он отвечает за содержательную часть [4] и отражает качество получаемой информации – сигнализирует о зашумленности (отсутствуют данные, которые не несут смысловой нагрузки в рамках заданной тематики). Следовательно, информация, содержащая данные, не удовлетворяющая поставленному критерию – незначима.

Также стоит учитывать, что Internet делает потенциально доступными огромные объемы ин-

формации и, тем самым, ставит новые проблемы эффективной работы с такими объектами. В ситуации «информационной перегрузки» особенно актуальными становятся автоматические методы работы с большими объемами информации [5].

Исходя из этого, можно предположить, что любая специализированная система, целью которой является извлечение значимой информации из определенного источника, должна заранее определить, с каким типом документа ей предстоит работать. А для этого системе необходимо скорректировать свой алгоритм работы автоматически под определенную структуру данных. При этом она должна разбить процесс обработки и извлечения знаний из источников информации на несколько этапов (рис. 1).

Как видно из рис. 1, специализированной системе необходимо выполнить ряд дополнительных действий, прежде чем перейти непосредственно к извлечению значимой информации из документа. Причем данные действия должны быть выполнены в четко определенной последовательности, в противном случае это может привести к сбою всей системы. Это оказывает негативное влияние на производительность почти любой системы с указанным выше принципом работы. В конечном счете, автоматическое определение типа документа увеличивает время поиска значимой информации для каждого исследуемого документа. Но если заранее определить его тип и выполнять анализ нескольких источников информации параллельно, то можно повысить производительность всей системы и тем самым уменьшить временные затраты на поиск необходимого документа (рис. 2)

Как видно из рис. 2, специализированная система за одну и ту же единицу времени может обработать в 3 раза больше документов при условии схожести алгоритмов и равном объеме извлекаемых данных. Хотя при этом область применения данной системы сужается из-за работы только с определенным типом документов.

В настоящие время существуют системы, способные работать как с документами любых типов, так и с одним, заранее определенным. Далее рассмотри некоторые из них.

2. Обзор существующих решений

Примером системы, которая способна работать с документами любых типов, является KIM – Semantic annotation platform [6]. Данная система отвечает за извлечение и обработку данных, получаемых из различных информационных источников.

Рис. 1. Этапы работы специализированной системы по извлечению информации из заданного документа любого типа

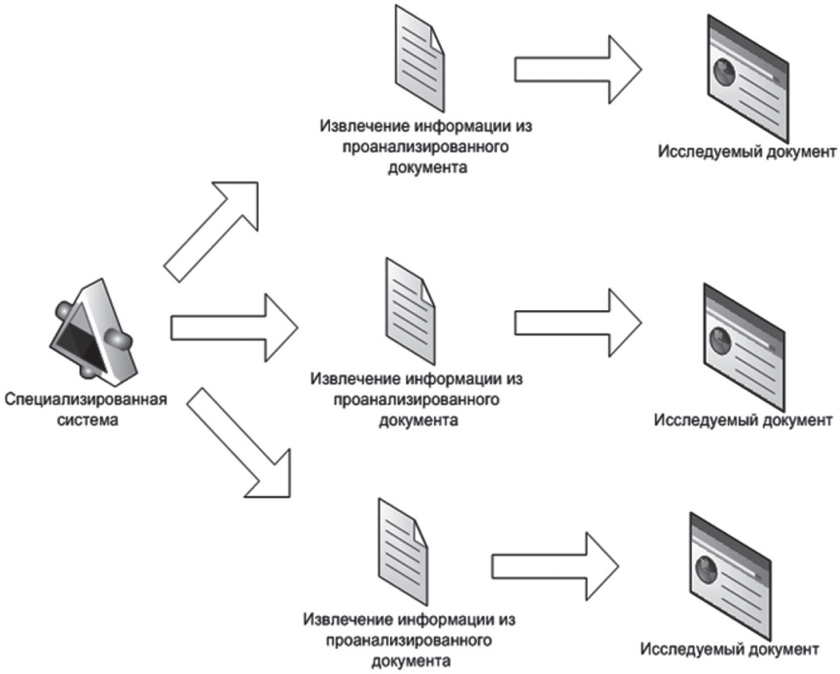

Рис. 2. Параллельная работа специализированной системы по извлечению информации из заданных документов заранее определенного типа

Она обеспечивает выполнение поставленных задач путем автоматического индексирования документов любого типа и построения их семантических аннотаций на основании проведенного анализа полученных данных. Для этого формируется база знаний, основанная на онтологии высшего порядка, в которой хранятся семантические аннотации в виде ключевых объектов (опорных слов) проиндексированных документов. Таким образом, в любом документе выделяются данные, которые соответствуют определенным классам, описанным в созданной онтологии. Каждый из классов, в свою очередь, делится на подклассы. Названия классов и подклассов соответствуют определенному общему термину, к которому можно отнести каждый из ключевых объектов (ключевой объект может соответствовать только одному классу или подклассу).

Все проаннотированные документы могут быть разделены на группы (на основании проанализированной meta-информации), каждая из которых соответствует определенной предметной области. Причем все данные в рамках одной группы связаны между собой. Это обеспечивает доступ ко всем ключевым объектам соответствующей тематики из документа любого типа, который был предварительно обработан данной системой.

К достоинствам KIM – Semantic annotation platform можно отнести:

1. Работу с любыми типами документов.

2. Взаимосвязь всех ключевых объектов различных документов, данные которых хранятся в базе знаний системы.

3. Возможность выполнения поиска документов по ключевым объектам.

К недостаткам KIM – Semantic annotation platform можно отнести:

1. Необходимость выполнения дополнительной обработки документа при создании семантических аннотаций.

2. Отсутствие проверки на зашумленность и повторяемость информации в исследуемых документах.

3. Неиспользование ключевых объектов для определения тематики документа. Ключевые объекты несут чисто информативный характер.

4. Отсутствие автоматизации составления логических правил, по которым ключевые объекты относят к какому-либо классу или подклассу, а также онтологии высшего порядка

5. Отсутствие возможности вносить в структуру онтологии высшего порядка изменения в процессе роботы системы.

Примером системы, которая способна работать с документами определенного типа, является программа, основанная на методе извлечения значимой информации из web-страниц путем их разделения на содержательную и навигационную часть. Данная система использует алгоритм, основанный на выделении повторяющихся фрагментов страниц одного сайта. Для этого на вход алгоритму подается директория с файлами, которая соответствует страницам одного сайта. После этого он анализирует данные файлы и выделяет в них повторяющиеся фрагменты, которые считаются навигационной частью. В зависимости от настроек, алгоритм либо удаляет навигационную часть из файла, либо выделяет навигационную часть специальными тегами. В свою очередь неповторяемые фрагменты относят к содержательной части, которая используется при информационном поиске документа, соответствующего формализованному запросу пользователя [4].

К достоинствам данной системы можно отнести:

1. Выполнение функций поиска данных, соответствующих запросу, сформированному пользователем, только в содержательной части web-документа.

2. Эффективную обработку информации на сайтах форумов, блогов, web-конференций, которые имеют стандартную структуру.

3. Возможность периодического мониторинга фиксированного списка сайтов.

К недостаткам данной системы можно отнести:

1. Необходимость выполнения дополнительной обработки web-документа при его разделении на содержательную и навигационную часть.

2. Низкая эффективность поиска web-документа, в случае если в его навигационной части содержится информация, релевантная сформированному запросу.

3. Отсутствие проверки на зашумленность содержательной части web-документа

4. Наличие случаев, в которых навигационную часть невозможно выявить или она выявлена неправильно на основе анализа совпадающих частей страниц.

5. Отсутствие функции лексического анализа содержательной части web-документа при поиске релевантных данных по сформированному запросу.

6. Работу только с одним типом документов.

Исходя из этого, можно сделать вывод, что независимо от количества поддерживаемых типов документов различными системами, они обладают рядом характерных недостатков. Главным образом они заключается в отсутствии возможности выявления и исключения шумов при извлечении данных из информационных источников.

Решением указанной проблемы посвящена данная статья.

3. Общие проблемы рассмотренных систем

Основываясь на приведенных выше исследованиях, было выявлено, что эффективного решения проблемы деления данных из любого информационного источника на значимую и незначимую

часть не предложено. В основном это вызвано наличием ряда следующих причин:

1. Большинство систем опираются на ключевые слова, которые могут иметь несколько значений и относится к разным тематикам, и поэтому возможно появление документов, не связанных со сформированным запросом.

2. Часто при анализе страницы web-документа исследуется только meta-данные, при этом другая информация не рассматривается, как следствие, возникают шумы в полученных данных.

3. В системах, направленных на извлечение информации из документов, отсутствуют критерии, характеризующие качество получаемой информации.

4. В процессе работы систем, направленных на извлечения информации, не формируются список «надежных» источников, данные из которых содержат наименьшее количество шумов. Это может отрицательно сказаться на эффективности работы всей системы в целом.

5. Большинство систем не выделяет значимую информацию из проанализированного источника и не хранит её в формате базы знаний. Тем самым данные системы сталкиваются с необходимостью повторной обработки данных во время появления новых поисковых запросов.

4. Постановка задачи

Исходя из изложенных выше причин, необходимо создать интеллектуальную систему, способную извлекать значимую информацию из документов с минимальным количеством шумов. Также она должна быть лишена основных недостатков и проблем, выявленных при рассмотрении схожих систем.

Для повышения эффективности будущей системы было принято решения обрабатывать динамические web-документы только с частично-структурированными данными (html-документами) в связи с тем, что она рассчитана на работу преимущественно с Internet ресурсами, где данный тип документов наиболее распространен.

5. Метод решения

На основании сформированной выше задачи предлагается рассмотреть модель работы системы извлечения значимой информации, которая способна повысить критерии качества получаемых данных. Данная модель отталкивается от предположения, что любой современный информационный web-документ можно разбить на различные структурные блоки. Каждый из этих блоков в свою очередь содержит определенные данные, посвященные заданной тематики, и выделен произвольным набором повторяемых html-тегов. В качестве примера одного из таких блоков возьмем часть html-кода, взятого из сайта футбольного клуб “Металлист” (http://www.metallist.kharkov.ua). Итак, данный блок имеет следующую структуру:

* <h2>МЕТАЛЛИСТ ОПРОС</h2> *

* Как, по Вашему мнению, завершится матч Металлист - Таврия? *

* голосовать *

Его границы были выделены по следующему принципу: ключевой html-тег, в данном случае это , не может быть частью другого тега. К примеру, тег <h2> является частью рассмотренного выше ключевого тега, следовательно, он не может быть началом другого структурного блока. Закрытие ключевого тега ( ) символизирует окончание описания данного структурного блока. Исключение составляют теги, отвечающие за формирование и описание общей структуры любой web-страницы.

Если учесть, что структуру современного динамического web-документа можно представить в форме ориентированного графа [7], корнями которого являются гипертекстовые ссылки, представленные в виде меню-навигации, то блок, который содержит ссылки на внешний источник или другую страницу из другого домена (за исключением ссылок встречаемых в обширных текстовых описаниях), можно принять за шум. А значит, они не должны учитываться при обработке web-страницы данной системой. Остальные информационные блоки анализируются и записываются в базу знаний в соответствии с критериями, которые были предварительно заданы пользователем.

Общий принцип модели системы, отвечающей за извлечения значимой информации из web-страницы, рассмотрен ниже (рис. 3).

Далее остановимся подробнее на критериях, которые предварительно задаются пользователем. В общем случае они могут выглядеть в форме простых пожеланий в формате предоставляемых данных. К примеру, пользователи могут сформировать списки любимых источников и обмениваться ими между собой в форме rdf-файлов (рис. 4).

Также в процессе работы данной системы осуществляется подсчет качества информации, которая содержится на текущей web-странице, по формуле:

$$I_q = I_{al} - I_b,$$

где I_q – процент значимой информации на текущей web-странице; I_{al} –процент всей информации на текущей web-странице (обычно равен 100%); I_b – процент шумов на текущей web-странице.

Основываясь на данной формуле, можно выстроить внутренний рейтинг приоритета обработ-

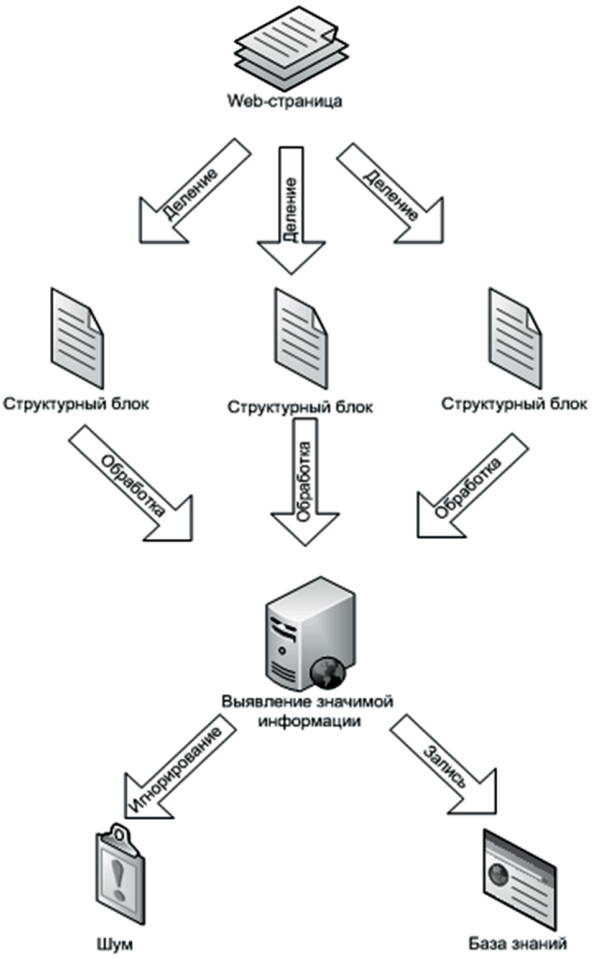

Рис. 3. Общая структурная схема модели системы, отвечающей за извлечение значимой информации из динамической web-страницы

ки информационных источников с наименьшим количеством шумов при условии схожести их тематик.

5. Анализ полученных результатов

В рамках рассмотренной модели системы, отвечающей за извлечения значимой информации из динамического web-документа, была сформирована база знаний сайта футбольного клуба “Металлист”. Каждый класс полученной онтологии соответствует пункту навигации данного сайта. Причем его название сформировано с использованием терминологический словаря онтологий по соответствующей предметной области [8].

Стоит отметить, что экземпляры каждого из этих классов хранят информацию, полученную при анализе web-страниц сайта футбольного клуба “Металлист” (рис. 5). Причем в этой онтологии данные могут быть представлены как текстом, так и картинками или видео. Также для каждой web-страницы формируются «белые» и «черные» списки, информация из которых автоматически считается значимой в первом случае или шумом, во втором.

Далее рассмотрим некоторые из основных полей, которые отображены на данном рисунке:

1. hasadress – хранит URL страницы, из которой были взяты данные;

2. hasname – хранит имя страницы, из которой были взяты данные;

3. haspictures – хранит URL картинок страницы, из которой были взяты данные;

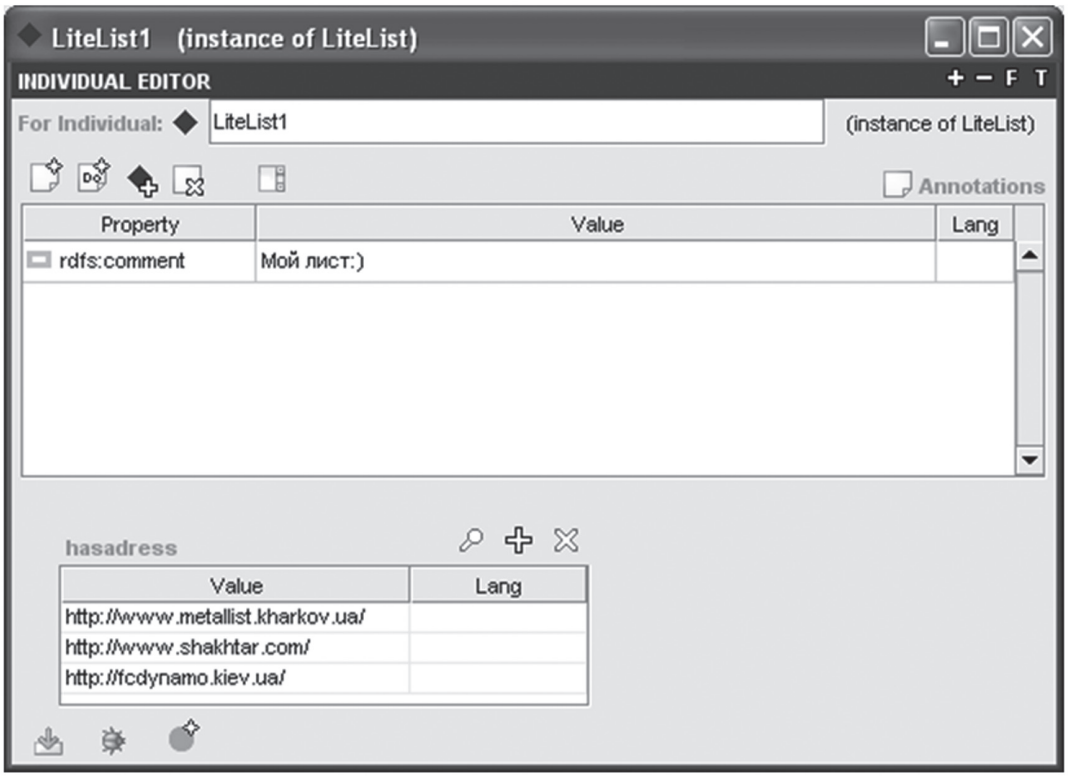

Рис. 4. Пример отображения списка любимых источников в редакторе онтологий Protйgй

4. hasLiteList – хранит URL страниц, которым доверяет система;

5. hasBlackList – хранит URL страниц, которым не доверяет система;

6. hasquality – хранит процент значимой информации от общего числа;

7. hasresorse – хранит значимую информацию, которая была взята из данной web-страницы (рис. 6).

В заключение можно сделать вывод, что полученная онтология позволяет сократить время поиска нужной информации за счет структуриро-

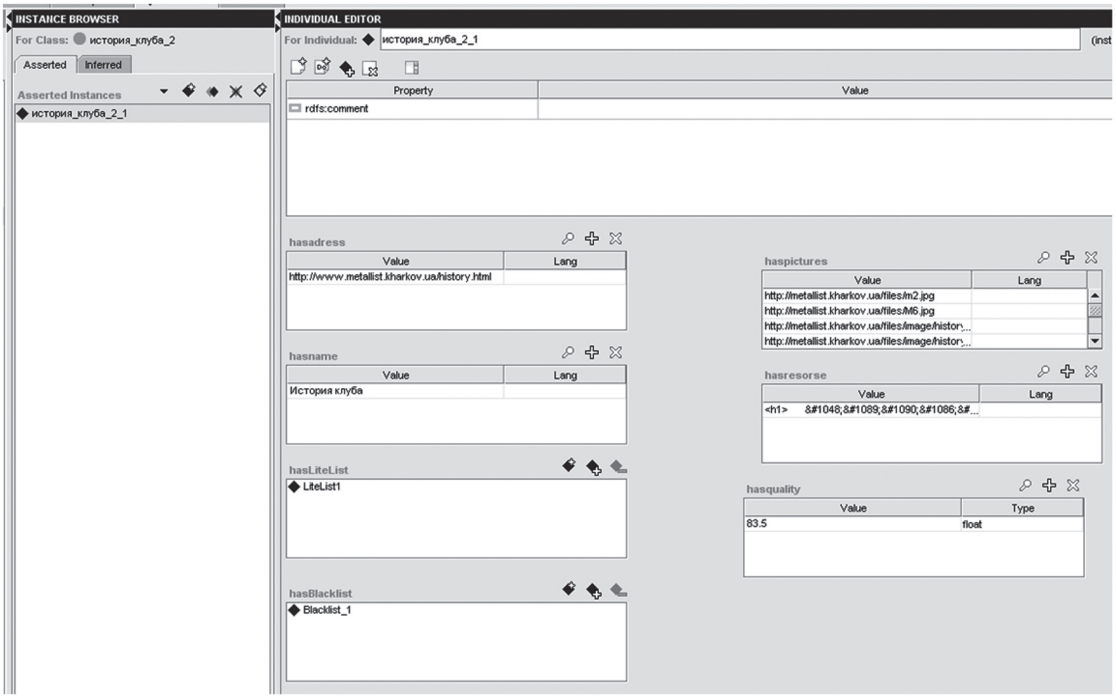

Рис. 5. Форма представление данных взятых из web-страницы сайта футбольного клуба “Металлист”

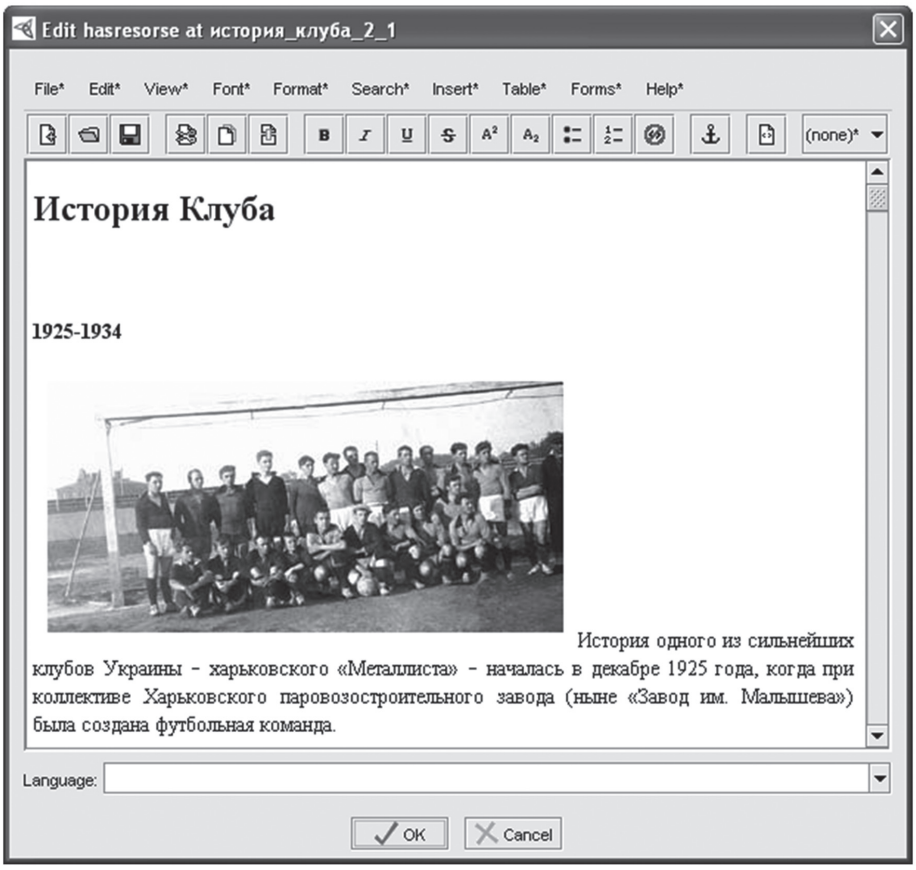

Рис. 6. Выделенная значимая информация из web-страницы (http://www.metallist.kharkov.ua/history.html)

вания источников данных. Причем в полученном результате будут отсутствовать шумы, что ускорит процесс его обработки и скажется на эффективности применения сформированной базы знаний в смежных системах.

Выводы

Полученная интеллектуальная модель системы, отвечающей за извлечения значимой информации из web-страниц, обладает рядом следующих преимуществ:

1. Извлекаемая информация проверяется на наличие шумов и повторяющихся данных, которые не записываются в базу знаний системы.

2. В процессе работы системы формируется белый список ссылок, которые должны быть проработаны в первую очередь, и черный список, ссылки из которого не анализируются, а полученные данные из них автоматически считаются шумом.

3. Хранение и обновления данных из проанализированных динамических web-документов осуществляется в формате единой базы знаний, через которую также выполняется поиск информации, релевантной сформированному запросу.

4. Для учета лексической взаимосвязи данных используется терминологический словарь онтологий.

5. Значимая информация в базе знаний разделяется за формой представления (картинки, текст, видео).

Подводя окончательную черту в описании модели системы, отвечающей за извлечения значимой информации из web-страницы, нужно отметить, что сформированная база знаний может использоваться при создании специализированных поисковых систем [8], а именно – в процессе анализа заданного web-документа, применяя и используя описанные выше особенности.

В последующих разработках планируется реализовать механизм проверки ссылок, встречаемых в обширных текстовых описаниях, а также применить описанную выше модель в рамках создание единой поисковой системы.

Список литературы: 1. *Chia-Hui, Ch.* A survey of Web Information Extraction [Text] / Ch. Chia-Hui, K. Mohammed, R.G. Moheb, F. S. Khaled. // IEEE Transactions on Knowledge and Data Engineering – 2006. №18/10. – С. 1411-1428. **2.** Информация [Электронный ресурс] / Википедия – интернет энциклопедия. – Режим доступа: URL: http://ru.wikipedia.org/ – 18.09.2011. **3.** *Беленький, А.* Текстомайнинг. Извлечение информации из неструктурированных текстов [Электронный ресурс] / А. Беленький // Журн. "КомпьютерПресс". – 2008. №10. Режим доступа: URL: http://www.compress.ru/article.aspx? id=19605&iid=905 – 18.09.2011. **4.** *Агеев, М. С.* Извлечение значимой информации из web-страниц для задач информационного поиска [Текст] / М. С. Агеев, И. В. Вершинников, Б. В. Добров // Интернет-математика 2005. Автоматическая обработка веб-данных. – М.:"Яndex", 2005. – С. 283-301. **5.** *Браславский, П.* Автоматическое реферирование веб-документов с учетом запроса [Текст] / П. Браславский, И. Колычев // Интернет-математика-2005. Автоматическая обработка веб-данных. – М. : "Яndex", 2005. - С. 485-501. **6.** *Popov, B.* KIM – Semantic Annotation Platform [Text] / B. Popov, A. Kiryakov, D. Manov, D. Ognyanoff, M. Goranov // Journal of Natural Language Engineering, №10/3-4. – С 375-392. **7.** *Ланде, Д.В.* Поиск знаний в Internet. Профессиональная работа [Текст] : пер. с англ. – М.: Издательский дом "Вильямс", 2005. – 272 с. **8.** *Почанский, О.М.* Модель построение адаптивных Web-страниц на основе интеллектуального анализа сети Internet [Текст] / О.М. Почанский // Журн. восточно-европейский журнал передовых технологий. – 2010. - № 4/7(46). – С. 66-69.

Поступила в редколлегию 19.09.2011

УДК 004.853

Витяг частково-структурованої (значущої) інформації з динамічних Web-документів / О.М. Почанський // Біоніка інтелекту: наук.-техн. журнал. – 2011. – № 3 (77). – С. 143-149.

В даній роботі описана модель інтелектуальної системи, що відповідає за вилучення значущої інформації з Web-сторінок. Це виконується шляхом поділу кожної сторінки аналізованого динамічного Web-документа на структурні блоки. Потім ці сторінки перевіряються на наявність шумів. А далі ті з них, які пройшли перевірку, зберігаються в базі знань. Результатом роботи системи є база знань, заповнена якісною інформацією (відсутні шуми) по заданій тематиці, яка може бути використана при створенні ефективних пошукових систем.

Іл. 6. Бібліогр.: 7 найм.

UDC 004.853

Removing partially-structured (significant) information from dynamic Web-documents / OM Pochansky // Bionics of Intelligense: Sci. Mag. – 2011. – № 3 (77). – P. 143-149.

The article describes the model of the intelligent system which is responsible for extracting meaningful information from Web-pages. Its main task is to divide each page of the analyzed dynamic Web-documents into different parts. Then they tested for the presence of noise, after that they saved into a knowledge base. The result of the system is the knowledge base that filled with quality information (without any noise), according to the chosen topic, which can be used to create effective search engines.

Fig. 6. Ref.: 7 items.