Схема автономного получения информации из новостных сообщений

Аннотация: В статье был приведен порядок автоматического извлечения полей из веб-страниц новостных сообщений. Суть алгоритма заключается в определении и отборе целевого узла DOM-модели (Document Object Model) HTML-документа статистическими методами группирования. Разработанный алгоритм отличается от существующих, главным образом, схемой работы, заключенной в две стадии. При этом модель признаков, необходимая для проведения классификации, имеет существенно расширенный вид. Благодаря предложенному методу извлечение информации из веб-страниц осуществляется по заданному образцу в автоматическом порядке. В результате была установлена процедура получения содержания, которая достаточно устойчива к видоизменению верстки и размещению целевых сведений на веб-странице.

Ключевые слова: новостные сообщения, графы сведений, схема получения информации, web-сайты

Введение

В интернете большая часть информационных агентств и средств массовой информации дают доступ к новостным сообщениям только лишь в виде HTML-страниц. Такой порядок удобен для конечного пользователя – текст имеет хорошее оформление и его удобно читать. Однако при этом обрабатывать сообщения в автономном режиме невозможно. Применение возможностей автоматического анализа базируется на структурированном представлении текста, который должен подразделяться на следующие элементы:

- заголовок новостного сообщения;
- дата и время публикации новости;
- непосредственно текстовое содержание.

Алгоритм кластеризации новостного потока был получен авторами экспериментальным путем вручную[1]. Недостаток такого решения заключается в плохой масштабируемости при значительном числе источников новостней.

На сегодняшний день для извлечения информации используют разработанные ручным способом принципы выделения составных частей, которые основываются на селекторах CSS (CSS – cascading style sheet) и XPath-выражениях[2]. Имеются также автоматические способы выделения содержания веб-страницы[3].

Метод, базирующийся на выделении составляющих и CSS-селекторов, достаточно трудный и затратный по времени. Каждый новостной источник получает собственный регламент получения информации. Специфика такого подхода заключается также в том, что при каком-либо изменении структуры веб-страницы разработанные ранее алгоритмы могут прекратить работу. В этой связи необходимо проверять, находятся ли они в рабочем состоянии.

---

[1] Казенников А.О. Анализ новостных потоков на основе информационного поиска и компьютерной лингвистики // Информатизация образования и науки № 4(16), 2012. – С. 155-164; Казенников А.О., Куракин Д.В., Трифонов Н.И. Гибридный алгоритм синтаксического разбора для системы анализа новостных потоков // Информатизация образования и науки № 1(13), 2012. – С. 90-97

[2] Sun F., Song D., Liao L. DOM Based Content Extraction via Text Density. In proc. Of SIGIR'2011, Beijing, China, 2011

[3] Kohlschutter C., Fankhauser P., Nejdl W. Boilerplate detection using shallow text features. In Proc. Of WSDM'10 pp. 441-450, 2010; Cai D., Yu S., Wen J., Ma W. Extracting content structure for web pages based on visual representation. In Proceedings of APWeb'03 pp. 406-417, 2003; Gottron D. Content code blurring: A new approach to content extraction. In Proc of DEXA'08, pp. 29-33, 2008; Vieira K., da Silva A., Pinto N. et al. A Fast and Robust Method for Web Page Template Detection and Removal. In Proc. Of CIKM'06, 2006; Chen L., Ye S., Li. X. Template Detection for lage scale search engines. In Proc. Of SAC'06, pp. 1094-1098, NY USA, 2006

Автоматические средства удаления разметки дают результаты при работе с простыми документами. Следует заметить, что современные веб-страницы имеют непростое строение – в них отображаются как навигационные элементы и блоки с рекламой, так и периодически подгружаемая информация, а также новые комментарии. Следовательно, после удаления разметки останутся лишь малоинформативные блоки (текст рекламных баннеров, навигационных составляющих и пр.).

Исследования последних лет, направленные на изучение особенностей автоматических методов анализа, были нацелены на определение основной части веб-ресурса[4]. Предложенные рекомендации позволяют решить лишь задачи выделения базиса, а для анализа новостного потока требуется извлечение именно структурированных данных.

В данной работе представлен метод, который объединил автоматический способ с алгоритмом выявления структурных составляющих новостных потоков. Главная особенность предложенного решения заключается в возможности получения отдельных элементов веб-страницы – заголовка, даты размещения, текстовой информации. Подобное решение может основательнее отбирать и распределять информацию по группам.

Формирование задачи

Автоматическое извлечение базиса новостей заключается в извлечении для конкретной страницы сведений по заранее утвержденной типологии:

- текстовой – суть заголовка и материал;
- временной – дата и время размещения.

Почти во всех случаях извлечение информации осуществляется при помощи DOM-моделии в виде представления HTML-страницы, которая представляет собой некое иерархическое отображение HTML-страницы с узлами в виде HTML-тегов и разделяющими их текстовый материал. Такая модель используется в качестве унифицированного средства для управления и внесения корректировок в состав страницы.

Хотя DOM-модель имеет хорошо организованную структуру, она все же не для структурного представления новости, так как HTML-страница имеет ориентир на отображение в браузере, а не на автономное исследование. Цель будет достигнута при получении интересующих сведений, точность которой дойдет до всех узлов DOM-дерева.

Возможные подходы

Ниже представлены базовые варианты автономного получения смысловой части:

- метод, выявляющий основную часть веб-страницы;
- выявление часто встречающихся элементов (шаблонов) страницы[5].

Основная масса нынешних подходов по определению и выделению основной части веб-страницы берет начало у статистических особенностей блоковых частей HTML-страницы. Построение данных алгоритмов заключается в следующем:

1) деление веб-страницы на основные смысловые блоки;

2) анализ полученных частей и выбор наиболее похожего на основной текст веб-страницы.

---

[4] Sun F., Song D., Liao L. DOM Based Content Extraction via Text Density. In proc. Of SIGIR'2011, Beijing, China, 2011; Kohlschutter C., Fankhauser P., Nejdl W. Boilerplate detection using shallow text features. In Proc. Of WSDM'10 pp. 441-450, 2010; Cai D., Yu S., Wen J., Ma W. Extracting content structure for web pages based on visual representation. In Proceedings of APWeb'03 pp. 406-417, 2003; Chen L., Ye S., Li. X. Template Detection for lage scale search engines. In Proc. Of SAC'06, pp. 1094-1098, NY USA, 2006

[5] Lin S., Ho J. Discovering informative content blocks from web documents. In Proc. Of SIGKDD'02, pp. 588-593, NY USA, 2002; Weninger T., Hsu W.H., Ma W. Learning block importance models form web pages. In Proc. Of WWW'04, pp. 971-980, NY USA, 2004; Bar-Yossef Z., Rajagopalan S. Template Detection via data mining and its applications. In proc. Of WWW'2002, pp. 580-591, 2002

В работе для автоматического определения составляющих был предложен порядок, предполагающий машинное обучение[6]. Согласно этому методу на основе тегов div, span, table, td выделяются основные блоки, после чего размечаются экспертами, а сформированная выборка применяется для обучения их классификатора на базе деревапринятия решений по принципу алгоритма C4.5[7]. Таким образом, создается дерево решений, подразделяя исходную обучающую выборку на 2 группы в зависимости от значения конкретного признака. На каждом этапе применяется тот признак, который для соответствующего множества примеров максимально точно разграничивает целевые классы. В ходе проведения исследования за основу брались признаки геометрических параметров блоков, а также ссылочная информация и структурные данные.

При методе частичной визуализации страницы, она преобразуется в дерево, узлами которого выступают внешне схожие сегменты[8]. В данном случае речь идет о поддереве DOM-представления документа, которое образует единое целое. Определение блока, который отображает главное содержание веб-страницы, происходит через оценивание разнородности его содержимого. В итоге ранжирование сегментов осуществляется как по смысловым характеристикам и пространственным (расположению, размеру) составляющим.

Метод извлечения содержания, базирующийся на плотности новостной информации, принимает во внимание количество текстовых символов на один тег[9]. Таким образом, предполагается, что основной блок веб-страницы содержит много текста и мало html-тегов. Такое предположение достаточно справедливо для новостей и блогов.

Иная картина наблюдается в блоках навигационного и рекламного характера. С этой целью можно рассмотреть усовершенствованный подход, который основывается на статистике ссылок[10]. Для анализа используются только содержательные блоки, остальные отсекаются при помощи порогового подхода. Смысл заключается в превышении показателя плотности определенного значения, что говорит об информативности блока и эффективности при установлении главного содержания веб-страницы. Однако могут возникнуть ошибки в той содержательной части, которая имеет большое количество тегов. К таким частям, в качестве примера, можно отнести перечень интересов участника социальной сети, места предыдущей работы, а также сложные структурированные поля, в которых элементы представлены отдельными тегами.

Существуют также принципиально отличающиеся способы выделения основной информации из страниц, например, подбор шаблонных (наиболее устойчивых) составляющих, суть которых заключается в следующем:

1) Основная масса современных сайтов применяет автоматическую генерацию веб-страниц. Такую картину можно наблюдать почти во всех сферах: сайты новостей, форумы, блоги, интернет-магазины и пр.

2) Предполагается, что с одного ресурса возможна генерация множества веб-страниц со схожим содержанием.

3) Подразумевается, что сформированный набор документов содержит лишь один уникальный шаблон для всех составляющих элементов. Следовательно, кластеризация совокупности по различным шаблонам не относится к постановке задачи выявления компонентов коллекции.

---

[6] Kohlschutter C., Fankhauser P., Nejdl W. Boilerplate detection using shallow text features. In Proc. Of WSDM'10 pp. 441-450, 2010

[7] Wu X., Kumar V., Quinlan J. Top 10 algorithms in data mining. Knowledge Information Systems Vol. 14, pp. 1-37, 2008.

[8] Cai D., Yu S., Wen J., Ma W. Extracting content structure for web pages based on visual representation. In Proceedings of APWeb'03 pp. 406-417, 2003

[9] Gottron D. Content code blurring: A new approach to content extraction. In Proc of DEXA'08, pp. 29-33, 2008

[10] Vieira K., da Silva A., Pinto N. et al. A Fast and Robust Method for Web Page Template Detection and Removal. In Proc. Of CIKM'06, 2006

Хотя удается достигнуть значительных результатов при определении основного содержания веб-страницы, представленный метод имеет и свои недостатки – он не фильтрует динамичные неинформативные части (ссылки на аналогичные новости, комментарии к теме и динамически внедренную рекламу).

Для поиска шаблонных компонентов применяют такие блоки, которые содержат однотипные функциональные элементы: ссылочный и навигационный блок, основной текст, меню и пр[11]. Документы серии подразделяются на части – для каждой из них, исходя из статистических характеристик, определяется гомогенность. Выделяют следующие критерии: количество ссылок в каждом разделе, их плотность на единицу текста, тега и др., а после этого части кластеризуются. Допускается, что если определенное количество блоков различных страниц оказалось в одной группе, то их можно считать шаблонными.

Метод подразделения страницы согласно тегам (например, table, div, span, td, dd), которые применяются для образования разных по смыслу информационных блоков, приводит к тому, что неинформационные части определяются подсчетом статистики слов в каждом блоке совокупности[12]. В противовес данному существует метод Site Style Tree, смысл которого заключается в формировании одного DOM-дерева[13]. Если определенный элемент имеется в SST, у него приумножается счетчик. Далее для каждого узла проводят оценку многообразия стилей содержания и отображения. Меньший показатель свидетельствует о том, что анализируемый компонент является шаблонным. В итоге можно определить неинформативные повторяющиеся составляющие.

Метод совместной процедуры выявления шаблонов и индексирования ведет к тому, что при индексировании страница подразделяется на части, далее осуществляется кластеризация исходя из стиля и позиции соответствующей части[14]. Подразумевается, что схожие группы различных документов – это шаблонная часть, подлежащая удалению.

Схема отбора составляющих структуры веб-страницы

Большая часть способов выделения информации основывается на машинном обучении, проводимом для определения главной части веб-страницы или шаблона. При идентификации их основного смысла зачастую применяют методы группировки. Что касается поиска шаблонов, то наиболее эффективными считаются методы кластеризации. Основная масса подходов применяет свойства верстки современных страниц. К таким можно отнести следующие:

- логическое разделение метода отображения вида и структуры веб-страницы;
- выделение в отдельную группу каждой информационной составляющей (теги span, div, table, li и др.);
- применение шаблонов стилей CSS.

Задача машинного обучения формально заключается в следующем:

1) Предположительно дано несколько размеченных страниц X, в которых представлены интересующие сведения.

2) Каждому DOM-элементу X соответствующей страницы из обучающего множества сопоставляется Y-метка, свидетельствующая о наличии в данном блоке интересующей информации.

3) В итоге образуется ряд пар (X, Y), пригодных для машинной обработки.

---

[11] Chen L., Ye S., Li. X. Template Detection for lage scale search engines. In Proc. Of SAC'06, pp. 1094-1098, NY USA, 2006

[12] Lin S., Ho J. Discovering informative content blocks from web documents. In Proc. Of SIGKDD'02, pp. 588-593, NY USA, 2002

[13] Weninger T., Hsu W.H., Ma W. Learning block importance models form web pages. In Proc. Of WWW'04, pp. 971-980, NY USA, 2004.

[14] Bar-Yossef Z., Rajagopalan S. Template Detection via data mining and its applications. In proc. Of WWW'2002, pp. 580-591, 2002

С целью классификации из DOM-элементов берутся следующие свойства[15]:

а) Свойства текста, скрытого в анализируемом блоке:

- длина;
- количество слов;
- средний размер слов;
- число предложений;
- средний размер предложений.

б) Свойства DOM-узла:

- название тега;
- перечень CSS-классов.

в) Расположение DOM-элемента в DOM-дереве:

- количество прямых потомков определенного элемента;
- длина расстояния до вершины DOM-дерева;
- количество родственных связей (потомки родителя);
- присутствие вложенных элементов в анализируемом DOM-элементе.

Главной отличительной особенностью представленного алгоритма считается его значительное расширение за счет характерных типов исследуемого компонента:

- биграммы, тригаммы и квадриграммы символов названий CSS-классов и идентификаторы;
- n-граммы прямого родителя компонента;
- комбинация групп, идентификатора и непосредственного родителя;
- средняя доля цифровых, буквенных, пунктуационных символов.

При помощи приведенных выше признаков можно извлекать как основное содержимое страницы, так и ее смысловые элементы. Потребность в расширенном способе возникает из-за того, что поставленная задача является более сложной, чем извлечение главного смыслового компонента веб-страницы. Для разрешения данной задачи можно прибегнуть к коллекции документов, тогда как для извлечения главного содержимого страницы такого рода набор документов отсутствует.

В качестве метода машинного обучения рекомендуется применять алгоритм машинно-опорных векторов или SVM (Support Vector Machine)[16]. Данный способ считается действенным при разрешении множества задач. К примеру, его применяли авторы в методе кластеризации новостей[17]. Еще одной характерной особенностью рассматриваемого алгоритма считается возможность задействования при выделении структурных компонентов двух этапов:

1) подбор осуществляется на основе обученного классификатора элементов (анализируются те характеристики, которые могли бы содержаться в данном компоненте);

2) удаление из выборки «лишних» компонентов посредством классификатора (образцы, состоящие из сведений, распределение которых значительно разнится от целевого).

Так удается получить более точные результаты отбора структурных компонентов. Приведенный способ состоит из обучающей и рабочей фаз. Первая включает в себя следующие шаги:

- маркировка многочисленной коллекции страниц, которая содержит интересующую информацию;

---

[15] Sun F., Song D., Liao L. DOM Based Content Extraction via Text Density. In proc. Of SIGIR'2011, Beijing, China, 2011; Chen L., Ye S., Li. X. Template Detection for lage scale search engines. In Proc. Of SAC'06, pp. 1094-1098, NY USA, 2006

[16] Wu X., Kumar V., Quinlan J. Top 10 algorithms in data mining. Knowledge Information Systems Vol. 14, pp. 1-37, 2008.

[17] Казенников А.О. Анализ новостных потоков на основе информационного поиска и компьютерной лингвистики // Информатизация образования и науки № 4(16), 2012. – С. 155-164

- образование обучающей совокупности на базе каждого DOM-узла соответствующей веб-страницы;

- создание классификатора DOM-компонентов, основываясь на имеющейся модели признаков.

В свою очередь, рабочая фаза включает в себя отбор из анализируемой страницы DOM-узлов, которые потенциально содержат нужные сведения. В общем случае отбираются лишь те, которые соответствуют блокам языка HTML: теги span, div, table td, h1-h6, li. Далее на базе классификатора, образованного на стадии обучения, осуществляется изъятие DOM-элемента, вероятнее всего несущих нужную информацию. На второй стадии проводят «отрицательную» выборку, смысл которой заключается в ликвидации из оставшихся компонентов наиболее неинформативных.

Предложенный алгоритм позволяет эффективно отбирать даже те компоненты, у которых в DOM-элементах имеются лишние данные. К примеру, в новостных сообщениях ресурса «Лента.ру» может включаться ссылка на материал по сюжету сообщения. Одностадийный способ, в свою очередь, может изъять лишь текст новости полностью, в том числе и подобную врезку. Разработанный способ полезен для определения и удаления нежелательной информации на втором этапе.

Оценивание алгоритма экспериментальным путем

С целью проверки действенности предложенного алгоритма вручную разметили по 2000 новостей, взятых из сайтов «Лента.ру», «Interfax», «NewsRu.COM» и «РИА Новости». В данной выборке разметке подлежали:

- название новостного сообщения;
- дата выхода новости;
- текст сообщения.

На данных с ресурса «Лента.ру» проводилась обучающая фаза метода. Далее на базе сайтов «Interfax», «NewsRu.com» и «РИА Новости» воплощалась рабочая – проводилась оценку таких характеристик, как полнота, точность и F-мера (гармоническое среднее от полноты и точности). Устойчивость метода обучения дополнительно оценивалась на данных сайта «РИА Новости», а результаты проверялись по оставшимся отобранным сведениям.

Эксперимент проводился в три этапа:

1) Установление влияния дополненной модели признаков на точность извлечения компонентов веб-страницы. Сформированный классификатор применялся лишь для извлечения целевых составляющих, но никак не для устранения лишних элементов.

2) Оценка эффекта, полученного от применения двустадийной схемы извлечения компонентов.

3) Оценка эффективности алгоритма выделения главных компонентов новостей при кластеризации информационных потоков.

В предложенном алгоритме кластеризации информационных потоков подразумевается, что новостные источники публикуют новости в пригодном для обработки виде[18]. Помимо этого, осуществлялась оценка роли каждой группы характеристик:

- текстовые свойства;
- свойства расположения DOM-узла;
- параметры DOM-узла;
- предложенное увеличение модели.

[18] Казенников А.О. Анализ новостных потоков на основе информационного поиска и компьютерной лингвистики // Информатизация образования и науки № 4(16), 2012. – С. 155-164; Казенников А.О., Куракин Д.В., Трифонов Н.И. Гибридный алгоритм синтаксического разбора для системы анализа новостных потоков // Информатизация образования и науки № 1(13), 2012. – С. 90-97

Результаты первой экспериментальной части приведены в таблице 1. Следовательно, можно сделать вывод, что расширение модели положительно сказывается на качестве выборки. Также нужно заметить, что ее эффективность не зависит от обучающей группы, так как она улучшала качество выборки в обоих случаях. F-мера для предложенного алгоритма существенно выше показателя, характерного для базовой модели (совокупность характеристик обозначается как «1+2+3»).

Таблица 1 – Результаты оценки зависимости качества отбора информации от модели признаков

Признаки	Источник данных	Точность	Полнота	F-мера
1	Лента.ру	0,881	0,892	0,885
1+2	Лента.ру	0,893	0,891	0,890
1+2+3	Лента.ру	0,912	0,923	0,915
1+2+3+4	Лента.ру	0,937	0,920	0,926
1	РИА Новости	0,898	0,901	0,895
1+2	РИА Новости	0,895	0,903	0,895
1+2+3	РИА Новости	0,914	0,921	0,915
1+2+3+4	РИА Новости	0,921	0,942	0,932

Итоги второго экспериментального этапа приведены в таблице 2. В процессе реализации данной серии экспериментов применялась модель признаков. Как и на первом этапе, определялось влияние двустадийной схемы при различных обучающих коллекциях. Как свидетельствуют данные из таблицы 2, она существенно повышает качество отбора и не зависит от обучающей коллекции. Такой результат достигается благодаря эффективной фильтрации.

Таблица 2 – Результаты оценки зависимости качества отбора информации от реализации двустадийной схемы

Схема	Источник данных	Точность	Полнота	F-мера
Одностадийная	Лента.ру	0,937	0,920	0,926
Двустадийная	Лента.ру	0,950	0,921	0,936
Одностадийная	РИА Новости	0,921	0,942	0,932
Двустадийная	РИА Новости	0,927	0,943	0,935

Сводную диаграмму результатов можно рассмотреть на рис. 1. Предложенная модель предоставляет возможность существенно повысить качество выборки.

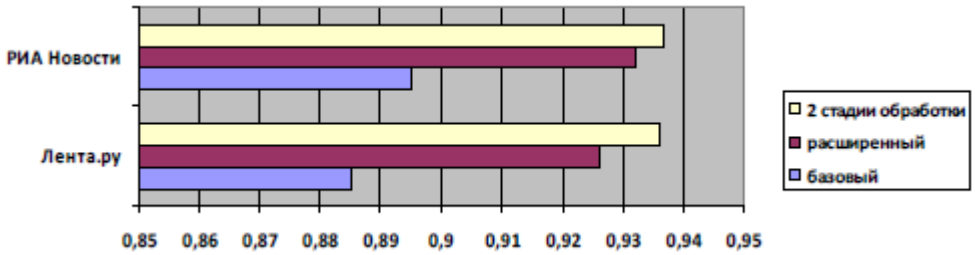

Рисунок 4 – Сводный график результатов экспериментов по группе сведений (F-мера)

Завершающий этап был проведен для анализа влияния предложенного алгоритма отбора компонентов веб-страницы на задачу кластеризации информационных потоков. Для корректности сопоставления результатов с работой применялась база «РИА Новости», так как для оценивания качества кластеризации был задействован материал ресурса «Лента.ру». Полученные результаты приведены в таблице 3. Разработанный метод дал результаты, сопоставимые с составлением правил для ручной выборки. Разница итогов по F-мере составляет 0,006 – это незначительная величина, которой при массовом добавлении новостных источников на практике можно пренебречь.

Таблица 3 – Результаты использования разработанного метода для кластеризации информационных потоков новостных сайтов

Метод	Точность	Полнота	F-мера
Ручной	0,912	0,891	0,901
Автономный	0,904	0,890	0,895

Выводы

Таким образом, полученные при эксперименте результаты свидетельствуют о том, что предложенный в работе алгоритм отбора информативных элементов из веб-страниц можно считать достаточно эффективным. С его помощью можно извлекать как главную информационную часть документа, так и отдельные структурные компоненты. Работа допускается как с текстовой информацией (название и текст новости), так и с датой публикации сообщения.

Эксперимент позволил оценить переносимость на другие источники результирующей процедуры отбора, обученной на одном из них. Данный метод показал хорошую адаптивность в пределах сходного материала.

В результате экспериментальной оценки предложенного алгоритма удалось установить взаимосвязь расширенной модели признаков и двустадийной схемы, применяемой для отбора информативных элементов. Предложенный метод позволяет значительно повысить качество выборки.

В работе также проводилось сравнение автоматического и ручного способа отбора структурных компонентов с целью кластеризации информационного потока. Полученные при автоматическом методе результаты сопоставимы с ручным. Таким образом, предложенный алгоритм можно применять для подключения источников новостей к системам анализа информационных сообщений.

Литература

1. Казенников А.О. Анализ новостных потоков на основе информационного поиска и компьютерной лингвистики // Информатизация образования и науки № 4(16), 2012. – С. 155-164
2. Казенников А.О., Куракин Д.В., Трифонов Н.И. Гибридный алгоритм синтаксического разбора для системы анализа новостных потоков // Информатизация образования и науки № 1(13), 2012. – С. 90-97
3. Sun F., Song D., Liao L. DOM Based Content Extraction via Text Density. In proc. Of SIGIR’2011, Beijing, China, 2011
4. Kohlschutter C., Fankhauser P., Nejdl W. Boilerplate detection using shallow text features. In Proc. Of WSDM’10 pp. 441-450, 2010
5. Cai D., Yu S., Wen J., Ma W. Extracting content structure for web pages based on visual representation. In Proceedings of APWeb’03 pp. 406-417, 2003
6. Gottron D. Content code blurring: A new approach to content extraction. In Proc of DEXA’08, pp. 29-33, 2008
7. Vieira K., da Silva A., Pinto N. et al. A Fast and Robust Method for Web Page Template Detection and Removal. In Proc. Of CIKM’06, 2006
8. Chen L., Ye S., Li. X. Template Detection for lage scale search engines. In Proc. Of SAC’06, pp. 1094-1098, NY USA, 2006
9. Lin S., Ho J. Discovering informative content blocks from web documents. In Proc. Of SIGKDD’02, pp. 588-593, NY USA, 2002
10. Weninger T., Hsu W.H., Ma W. Learning block importance models form web pages. In Proc. Of WWW’04, pp. 971-980, NY USA, 2004.
11. Bar-Yossef Z., Rajagopalan S. Template Detection via data mining and its applications. In proc. Of WWW’2002, pp. 580-591, 2002

12. Wu X., Kumar V., Quinlan J. Top 10 algorithms in data mining. Knowledge Information Systems Vol. 14, pp. 1-37, 2008.