УДК 004.6

И.А. Черенков

Национальный технический университет «ХПИ», Харьков

АВТОМАТИЧЕСКИЙ ПОИСК ДАННЫХ ИЗ НОВОСТЕЙ НА ПРИМЕРЕ РЫНКА ПОЛИМЕРОВ

В статье рассматриваются вопросы автоматического извлечения данных из текстовой новости. Рассмотрен процесс перехода от поверхностных текстов к глубинным. Описаны особенности хранения информации в новости. Выделены проблемы извлечения данных из новостных потоков в интернете. Предложен подход по автоматической добыче данных из текстовой новости.

Ключевые слова: *добыча данных, новостной поток, структура новости, извлечение данных из новости.*

Введение

Интернет как источник новостей для автоматического ценового прогнозирования на данный момент не имеет аналогов. Такое положение достигается за счёт большой оперативности появления новостей, доступности, всеобъемлемости. Однако, существует ряд проблем, затрудняющих использование интернета как источника новостей. К таким проблемам можно отнести недостаточную надёжность новостей, чрезмерную избыточность потока новостей. При автоматической добыче данных из новостного потока круг проблем расширяется: слабая структурированность новостей, представление новости в текстовом формате (plain text) на разговорном языке, и, как следствие, сложность в извлечении незашумлённых данных. Таким образом, актуальна задача автоматической обработки новостей, которая не является тривиальной и требует разработки собственного подхода.

Постановка задачи. Выработаем подход для автоматического извлечения данных из интернет-новости текстового формата на примере рынка полимеров.

Подход

Обособленная новость является текстом, построенным на основе используемой в общении грамматики (разговорный язык).

С точки зрения грамматик текстовую новость можно представить в виде следующей иерархии моделей (рис. 1).

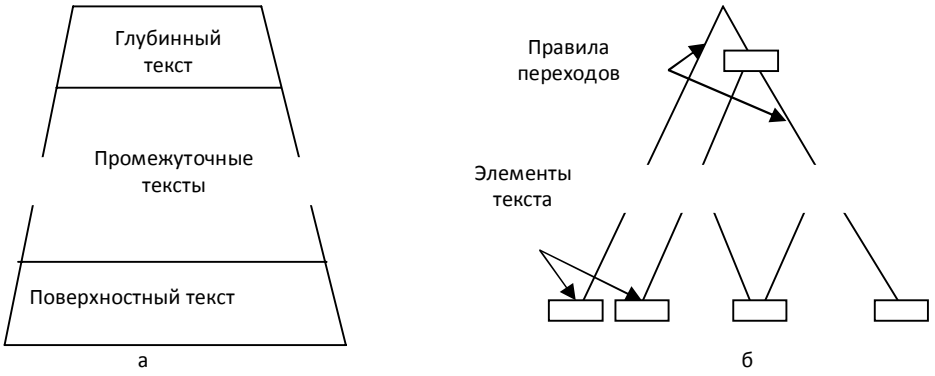

Рис. 1. Модель текста

Читаемая новость представляется в виде поверхностного текста рис. 1, а, а задача извлечения данных заключается в переходе от поверхностного к глубинному тексту, в котором заключён основной смысл. Изображённый на рисунке блок промежуточных текстов соответствует переходам на основе используемой грамматики и её правил от глубинного текста к поверхностному – рис 1, б. Вопросу формирования поверхностных текстов из некоторого глубинного уделено много внимания в работах Холмского, Мельчука [1, 2].

В общем случае у человека анализ новости как текста от внешнего уровня к глубинному проходит следующие этапы (рис. 2). Вначале происходит прочтение новости, т.е. происходит осознание того, что перед нами новость. Далее в зависимости от тематики активизируется соответствующая модель для анализа данных и непосредственно извлечение данных, на основе которых обновляется модель предметной области, для которой поступила новость [3]. Именно данные из обновлённой модели предметной области используются в дальнейшем при принятии решений, т.е. данные новости во взаимосвязи с другими данными предметной области.

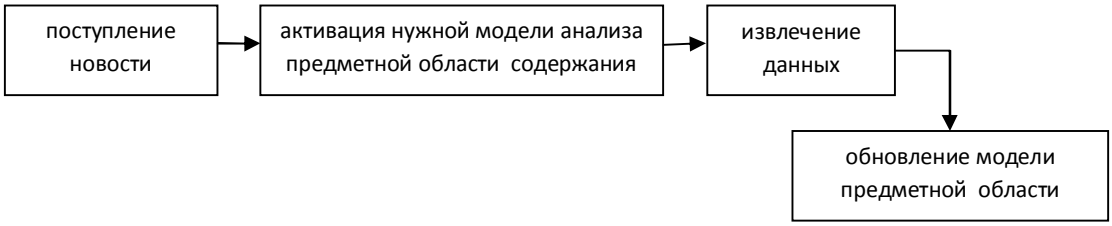

Рис. 2. Этапы анализа текста

Можно утверждать, что нет эффективного на данный момент общего подхода как по созданию поверхностных текстов, так и по восстановлению глубинного текста, например, ни один автоматический переводчик не переводит без абсурдных ошибок. Основная причина этому заключается в том, что для больших предметных областей, как в случае общего перевода, не существует детальной модели предметной области, способной составить альтернативу человеческому пониманию мира. Хотя в общем случае эффективное извлечение данных невозможно, но для более узкой предметной области, такой как новостной поток конкретной отрасли, создать такую модель можно. В простейшем случае модель анализа предметной области может быть представлена в виде графов с вершинами-объектами и переходами-действиями над объектами. Так для анализа новостных потоков вершины соответствуют категориям новостей. В общем случае – это бизнес правила модели предметной области. На примере полимерных новостей такая модель в виде графа будет выглядеть следующим образом (рис. 3).

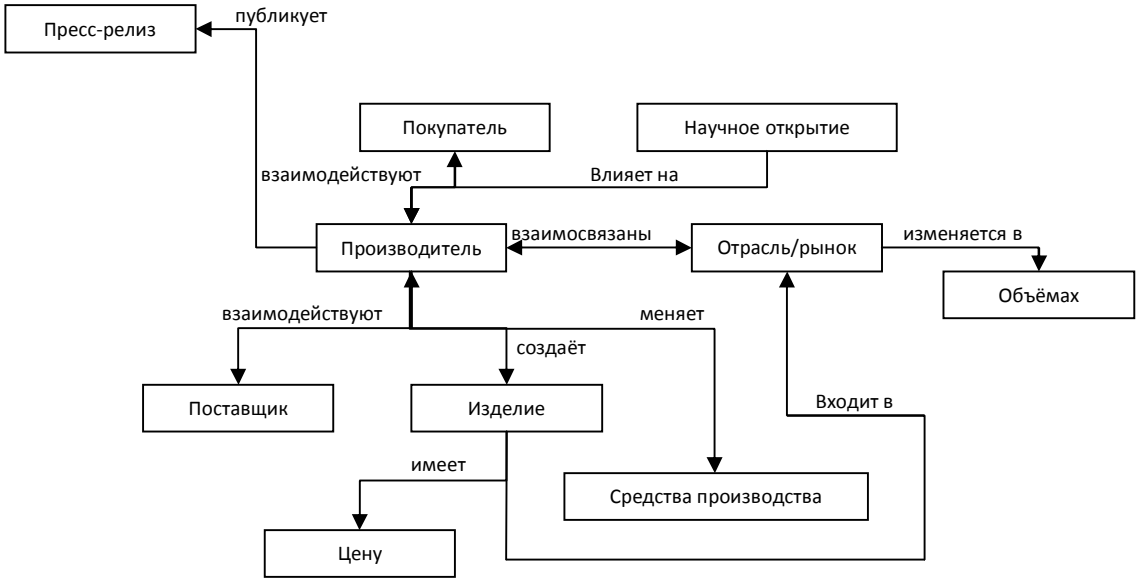

Рис. 3. Модель анализа текста для полимерных новостей

Основное отличие модели предметной области от модели анализа в том, что глубинный текст (смысл) выражается в элементах модели предметной области и она, соответственно, может быть использована для хранения данных о событии, в то время как модель анализа используется для отражения логики преобразований от глубинного текста к поверхностному. В целом модель анализа есть визуальное представление правил грамматики, где каждой взаимосвязанной паре объектов на рис. 3 соответствует своя детализированная модель анализа на основе конкретного набора правил грамматики.

Таким образом, для успешного решения задачи по автоматизации извлечения данных из новостного потока требуется исследование правил грамматик для всех пар объектов модели анализа. Пример модели предметной области для полимерной отрасли представлен на рис. 4.

Рассмотрим основные проблемы извлечения данных как из новости, так и новостного потока в целом:

– Слабая структурированность и текстовый формат. Новость как таковая является описанием автора в текстовом виде некоторого события, что обусловливает субъективный фактор в новости. При этом отсутствуют общепринятые стандарты написания (структуры), глубины новости. Вследствие чего новость представляет информационный «суп», включающий важную информацию о непосредственно возникшем событии и прочую информацию, значительно зашумляющую, а иногда и искажающую новость как информационную единицу. В результате извлечённая информация из новости является малообъективной и недостоверной.

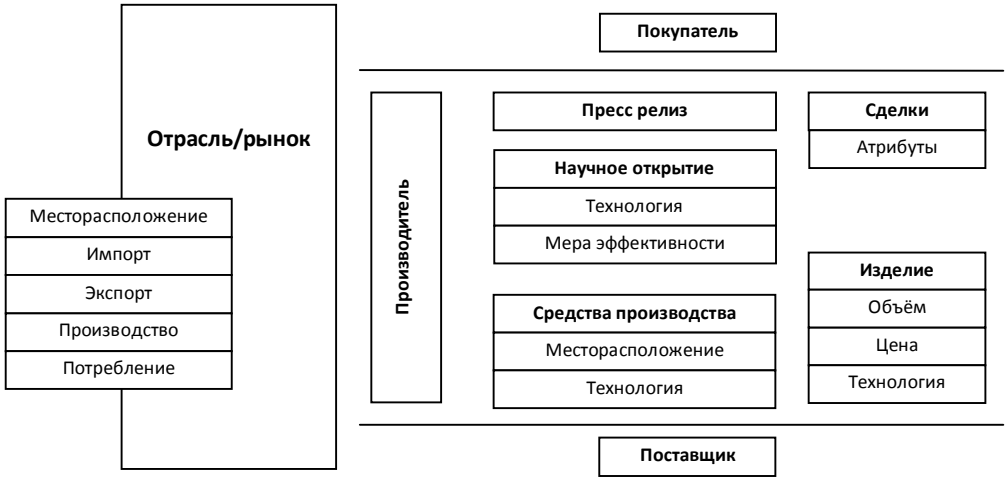

Рис. 4. Модель предметной области

Основной риск связан с ценой ошибки принятого решения на основе неправильно извлечённой информации из новости.

– Малоинформативные новости. К данной категории можно отнести те новости, что не являются значимыми, т.е. они описывают какое-то событие, однако их влияние на общую картину является незначительным, как следствие при автоматическом рассмотрении такой новости она может исказить реальную картину происходящего. Крайним случаем малоинформативной новости можно считать полностью искажённую новость. Как следствие понижается надёжность источника новости. В связи с этим возникает необходимость использования различных источников.

Однако, при использовании различных источников новостных потоков возникает другой ряд проблем, актуальных в первую очередь для автоматической обработки новостей. Этот круг проблем связан с восстановлением общей картинный событий, их последовательности и важности.

Дубликаты. Новости, описывающие важные события, будут дублироваться в различных источниках и будут характеризовать одно и то же событие. Сложность заключается в однозначной идентификации события, что наряду с вышеописанными проблемами слабой структурированности новостей составляет проблему. Основная ошибка – воссоздание дубликатов одного и того же события.

Проблема сюжетных (логических) цепочек новостей напрямую связана с проблемой дубликатов. Новости лишь описывают событие и нет взаимнооднозначного соответствия, как следствие множество новостей, не являющихся дубликатами, могут формировать некоторую сюжетную линию события, однако такие новости будут зачастую избыточными, после появления первой новости.

Простейшим решением вышеописанных проблем было бы введение стандартизации представления новости и описания события. В наиболее общем случае это мог бы быть некоторый искусственный язык передачи информации – универсальный язык. Однако, лаг, связанный с внедрением такого стандарта, и отсутствие механизмов контроля вместе с малой заинтересованностью в такой стандартизации делают этот подход нереализуемым. Таким образом, необходим поиск вариантов решений при предположении неизменности используемого языка.

Для разрешения проблемы слабой структурированности новости, т.е. задачи первичного извлечения информации, проанализируем новость как контейнер информации. Новость – это носитель информации о событии. По правилам журналистики [4] новость должна иметь следующую структуру (рис. 5). Однако на практике такая иерархия выполняется далеко не всегда, что в совокупности со сложностью автоматического распознавания текста делает невозможным использование исключительно такого шаблона для анализа всей новости целиком.

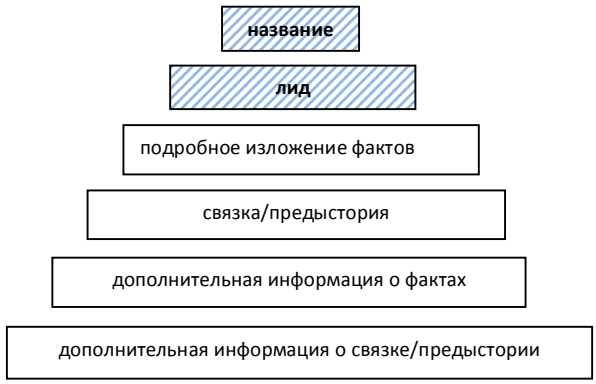

Рис. 5. Структура новости

Но если ставить целью получение необходимой и достаточной информации о событии, то целесообразным будут лишь вершины иерархии названия новости и лида, которые выдерживаются почти всегда. Лид – это начало статьи, по правилам, это должно быть простое и ясное высказывание о самом важном. В общем случае для новости лид содержит такие информационные элементы: «кто», «что», «когда», «где», «почему»,

«как». Эти же элементы могут отображаться и в названии. Структурированная таким образом информация будет необходимой и достаточной для принятия решения. Проблемы при таком анализе связаны с тем, что новость есть всегда личное изложение автора и обладает присущим ему стилем речи. Помимо этого новость не является абсолютно абстрактным носителем информации, т.е. в общем случае, как продукт, она должна быть ориентирована на её потребителей-читателей, а значит должна привлекать их внимание, в результате «голая» информативность новости обрастает различными подходами изложения и психологическими уловками для привлечения внимания, затрудняя тем самым извлечение данных событии, если не искажая их. Для украино- и русскоязычных источников новостей о полимерах в силу узкой специализированности информационного потока искажений, как в случае с политическими или социальными новостями, значительно меньше, однако они есть. Очевидно, что первоочередная задача заключается в извлечении элементов лида и названия. Наиболее общий подход – это использование лингвистических моделей текста, которые в совокупности со словарями конкретной предметной области, в нашем случае полимеров, смогут дать достаточно высокое качество извлечения информации.

Таким образом, для автоматической добычи данных из текстовой новости интернет-новостного потока необходимо следующее:

1. Формализация модели анализа предметной области для конкретного рынка.

2. Формализация правил взаимодействия объектов внутри модели. Данные правила отражают логику формирования глубинных текстов.

3. Связать правила формирования глубинных текстов с синтаксическими структурами названия и лида новости.

Необходимо отметить, что вопросы, связанные с техническим аспектом извлечения данных из новости в интернете, как сбор новостей, их парсинг, не являются объектом исследования данной статьи.

Вместе с тем, в новости остаётся достаточно большой пласт данных – это подробные сведения о событии, идущие после лида. Эти сведения позволили бы однозначно идентифицировать событие и его репрезентацию в других новостях. Однако, здесь практически полностью отсутствует структура данных, что делает незашумлённое извлечение данных практически невозможным. Одним из вариантов решения может быть введение дополнительных словарей слов-связок для синтаксических шаблонов дополнительных объяснений о событии и анализе по ключевым словам, осуществляя поисков паттернов, аналогичных поиску по лиду. Однако, эффективность такого подхода будет небольшой, в силу отсутствия структуры и непредсказуемой логики изложения.

Выводы

1. Автоматическое извлечение данных из новости возможно посредством использования правил переходов от поверхностного текста к глубинному на основе моделей анализа предметных областей. Выработанный подход применим только для новостного потока конкретного рынка, универсального подхода не существует.

2. Слабая структурированность новости представляет проблему для добычи данных лишь в том случае, если извлекаются данные из всей новости целиком. Необходимым и достаточным для добычи данных о событии является анализ элементов названия и лида новости.

3. Выработанный подход позволяет получить необходимые данные о событии высокого качества, достаточные для принятия решений на их основе.

Список литературы

1. Мельчук И.А. Опыт теории лингвистических моделей смысл-текст. Семантика. Синтаксис / И.А. Мельчук. – М.: Высш. шк., 1999. – 345 с.

2. Хомский Н. Аспекты теории синтаксиса / Н. Хомский. – Изд-тво БГК им. И.А. Бодуэна Де Куртенэ, 1999. – 258 с.

3. Хомский Н. Язык и мышление. Язык и проблемы знания / Н. Хомский. – Изд-во БГК им. И.А. Бодуэна Де Куртенэ, 1999. – 254 с.

4. Ким М.Н. Новостная журналистика. Базовый курс / М.Н. Ким. – Изд-во Михайлова В.А., 2005. – 352 с.

Поступила в редколлегию 21.09.2011

Рецензент: д-р техн. наук, проф. Е.Л. Пиротти, Национальный технический университет «ХПИ», Харьков.

АВТОМАТИЧНИЙ ПОШУК ДАНИХ З НОВИН НА ПРИКЛАДІ РИНКУ ПОЛІМЕРІВ

І.О. Черенков

У статті розглядаються питання автоматичного видобутку даних з текстової новини. Розглянуто процес переходу від поверхневих текстів до глибинних. Описано особливості зберігання інформації в новині. Виділено проблеми отримання даних з новинних потоків в інтернеті. Запропоновано підхід з автоматичного видобутку даних з текстової новини.

Ключові слова: *видобуток даних, новинний потік, структура новини, вилучення даних з новини.*

DATA MINING OF THE NEWS ON THE EXAMPLE OF POLYMERS MARKET

I.A. Cherenkov

Subject of this article is news data mining. Following aspect are being reviewed in the article: the process of transition from text to semantics; key problems of stored information in the news text body; problems of extracting data from the news stream in the Internet. An approach for data mining of textual news was proposed.

Keywords: *data mining, news stream, the structure of news, data extraction from news.*