УДК 004

ИНТЕЛЛЕКТУАЛЬНЫЕ АГЕНТЫ ДЛЯ СОЗДАНИЯ ПСЕВДО КОНТЕНТА ВЕБ-САЙТОВ

М.В. Збитнева

ХНУРЭ, г. Харьков, Украина, e-mail mayazbt@yandex.ru.

В настоящий момент популярным направлением является автоматическая генерация естественно-языковых текстов, результаты которой находят все большее применение в различных областях. Приведен анализ способов создания уникального контента, построены и расширены марковские модели, построен метод формирования псведо-контента, определена архитектура интеллектуального агента.

ИНТЕЛЛЕКТУАЛЬНЫЙ АГЕНТ, МАРКОВСКИЕ МОДЕЛИ, ГЕНЕРАЦИЯ КОНТЕНТА, СОЗДАНИЕ КОНТЕНТА, ВЕБ-САЙТ

Введение

В настоящее время одним из способов, влияющих на скорость заработка при помощи Web, является использование дорвеев [1, 2]. Волнует вопрос, сколько сайт будет находиться в поисковой системе и сколько на этом можно будет заработать. Дорвей (doorway) – это набор веб-страниц, созданных под конкретный поисковый запрос, для поддержания высокого рейтинга в поисковых системах. В дорвеях используется большое количество ключевых слов и учитываются различные факторы, влияющие на ранжирование. Формирование дорвеев происходит на основе множества ключевых слов и текста тематических статей, сайтов. Пользователи, попавшие на такой сайт, часто перенаправляются на другой сайт. Одним из ключевых моментов в разработке таких сайтов является генерация контента.

В работе [3] представлено описание методов реферирования. В работе [4] идет речь о генерации кода. В работе [5] – вопросы синтеза текстов на естественном языке. Целью данного исследования является анализ способов генерации контента. В качестве задачи исследования выбрано определение архитектуры интеллектуального агента для автоматического создания псевдо-контента веб-сайта, построение марковской модели и создание метода по данной модели.

Сайт состоит из контента и программного кода. Контент – это та часть, которую практически невозможно защитить. Существующие способы защиты контента обеспечивают маленькую вероятность. К ним относятся:

– публикация в открытой печати;
– регистрация авторских прав;
– цифровые водяные знаки;
– использование средств отслеживания плагиата.

Контент представляет собой информационное наполнение сайта: текст, графические, аудио и видео файлы. Далее под контентом будет пониматься только текстовая масса.

Содержимым дорвеев является псевдо контент. Создание псевдо контента подразумевает автоматическое создание контента сайта на естественном языке.

Наряду с дорвеями широкое применение нашел клоакинг. Клоакинг — это методика отображения роботом поисковой системы оптимизированной страницы, в отличие от страницы, которую увидит обычный пользователь.

1. Клоакинг

Клоакинг реализуется с помощью программ и скриптов, выполняющихся на стороне веб-сервера. Серверные скрипты формируют выходные данные в зависимости от изменяющихся параметров: параметров в адресе запроса, доступных системных переменных веб-сервера и окружения, некоторых параметров запроса. По ряду данных можно определить от кого исходит запрос — от робота или обычного пользователя и сформировать для каждого отдельную результирующую страницу.

Существует два основных метода клоакинга:

– метод определения IP;
– метод определения UserName.

Метод определения IP. IP адрес идентифицирует подключение к Интернет. Пауки поисковых систем также имеют свои IP. Причем каждый паук имеет свой фиксированный IP адрес. Метод сводится к тому, что просто нужно определить IP посетителя, сравнить этот IP с базой данных (в которой записаны IP пауков поисковых систем) и сделать вывод посетитель это или паук. Если это посетитель, то показываем ему нормальную страницу. Если это паук - показываем ему оптимизированную страницу. Причем при создании оптимизированной страницы не нужно оглядываться на благозвучие текста, дизайн и тому подобное, потому что никто, кроме паука, эту страницу никогда не увидит.

Недостатки: Нужно иметь большую базу данных IP пауков, которую нужно периодически обновлять. Это достаточно дорогое удовольствие.

Преимущества: При хорошей базе данных вероятность обнаружения того, что используется клоакинг, крайне мала. Следовательно, методы оптимизации останутся не известными.

Метод определения UserName. Суть этого метода та же, что и предыдущего. Только определять нужно не IP, а UserName. Последовательность

действий такая же: определяем, сравниваем, показываем нужную страницу.

Недостатки: UserName можно подделать и “прикинуться” поисковым пауком. И таким образом узнать все секреты оптимизации.

Преимущества: Можно без проблем получить базу данных UserName пауков. То есть простота реализации.

Оптимальное решение

Самый надежный способ реализации это совмещение двух методов. То есть определение и IP и UserName. И чтобы не было проблем, в случае если даже IP сходится, а UserName - нет, то показывать нормальную страницу.

Таким образом, одна из сфер применения генерации контента – это использование клоакинга со стороны сервера, поскольку робот не учитывает качество текстовой массы.

2. Способы создания уникального контента

Способы создания уникального контента:

– написанный вручную:

а) копирайтинг – создание уникального текста с целью продвижения на рынке товаров и услуг; самостоятельное написание статей для сайтов, блогов; покупка готового уникального контента на текстовых биржах;

б) рерайтинг – переписывание статьи своими словами с сохранением смысла для размещения на сайтах, публикация в СМИ;

в) перевод контента.

– сгенерированный контент:

а) копипастинг:

– составление шаблона предложений;

– заполнение шаблона из множества контекста.

б) уникализация текста:

– добавление опечаток;

– замена русских букв на латинские;

– смена предложений и абзацев местами.

в) синонимизация:

– подбор синонимов;

– перемещение слов в предложении;

– перемещение предложений.

г) компьютерная лингвистика.

Существуют также способы автоматической генерации форматирования текста и программного кода [4].

3. Способы автоматической генерации текста

Автоматическая генерация текста предполагает использование программного обеспечения, которое создает последовательности символов, как правило, лишенные смысла; позволяет создать быстро и уникально по правилам русского языка текст на заданную тему. Сгенерированный текст содержит существительные, прилагательные, глаголы и наречия.

При этом соблюдается естественное разнообразие и грамматически правильное согласование падежей, лиц, чисел, времен, форм, родов и так далее. Тексты могут проходить проверку орфографии в Word. Наличествуют различные знаки препинания, предлоги и союзы. Все предложения начинаются только с заглавной буквы, присутствуют сложные предложения.

Присутствует обработка слов по сочетаемости – между существительными, существительными с глаголами, существительными с прилагательными и глаголов с наречиями.

При генерации текста на определенную тему подбор слов осуществляется так, что слова в данной области будут встречаться намного чаще, чем остальные.

При автоматической генерации сложно добиться естественного языка, можно получить грамматически правильно читаемый текст без особого смысла, насыщенный разнообразными словами, иногда со смешными фразами и целыми фрагментами текста.

Автоматическая генерация текста также является одним из направлений компьютерной лингвистики [5], наряду с анализом текста, пониманием текстов, оживлением текстов, моделями коммуникации. Для синтеза текста могут использоваться актанты действий. С каждым действием связан набор объектов и характеристик. Например с действием «ехать» связаны субъект, который совершает движение, начало и конец движения и так далее. Все эти данные объединяются в структуры – фреймы. Таким образом, процесс синтеза текста выглядит следующим образом:

– генерация нужной последовательности глаголов;

– заполнение актантных структур, создание глубинной структуры предложений;

– связь структур по общим действующим субъектам и объектам;

– создание синтаксически правильных структур предложений.

Одной из областей применения являются системы автоматического реферирования [3], которые формируют краткое изложение либо путем выделения фрагментов информационного наполнения и последующего их соединения, либо методом генерации текста на основании знаний об оригинале. Здесь используется метод линейных весовых коэффициентов. Веса назначаются для блока текста: учитывают расположение блока в оригинале, частоту появления в тексте, частоту использования в ключевых предложениях, показатели статистической значимости. Весовой коэффициент расположения зависит от начала, конца, середины текста, его присутствия в ключевых разделах: введение, заключение. Статистическая важность вычисляется на основании данных, полученных в результате анализа автоматической индексации. Здесь учитывается появление термина в заголовке, колонтиту-

ле, первом параграфе, пользовательском профиле запроса. Данный метод не учитывает связи между предложениями. В ряде подходов создается специальное окно для предыдущего предложения реферата, с помощью которого можно определить наличие смыслового разрыва или «висящего» слова. В других случаях предложения, содержащие «висящие» слова, исключаются из реферата, либо делаются попытки разрешения ссылок, которые предполагают такие слова или их сверки путем краткого лингвистического анализа. При таком подходе степень сжатия уменьшается, так как в реферат привносится посторонняя информация. Кроме того, когда основной реферат уже сформирован, трудно восстановить исходный процент сжатия.

Следующий метод основан на подборе выдержек. Здесь необходимы грамматики, словари синтаксического разбора, онтологии. Данный метод использует традиционный лингвистический метод синтаксического разбора предложений. В этом методе применяется семантическая информация для аннотирования деревьев разбора.

Процедуры сравнения манипулируют непосредственно деревьями с целью удаления и перегруппировки частей, например путем сокращения ветвей на основании некоторых структурных критериев, таких как скобки или встроенные условные или подчиненные предложения. После такой процедуры дерево разбора существенно упрощается, становясь, по существу, структурной «выжимкой» исходного текста.

Данный подход относится к системам искусственного интеллекта и опирается на понимание естественного языка. Синтаксический разбор также входит составной частью в такой метод анализа, но деревья разбора в этом случае не порождаются. Напротив, формируются концептуальные репрезентативные структуры всей исходной информации, которые аккумулируются в текстовой базе знаний. В качестве структур могут быть использованы формулы логики предикатов или такие представления, как семантическая сеть или набор фреймов. Примером может служить шаблон банковских транзакций (заранее определенное событие), в котором перечисляются организации и лица, принимающие в нем участие, дата, объем перечисляемых средств, тип транзакции и так далее.

4. Интеллектуальные агенты для создания псевдо контента веб-сайтов

Одним из распространенных способов создания динамических приложений, меняющих свое поведение или состояние в режиме реального времени, являются интеллектуальные агенты. Из наиболее известных архитектур интеллектуальных агентов [6] для данной задачи выбрана архитектура агента, основанного на модели. Модель представляет собой расширенную марковскую модель, которая описывается ниже. Марковская модель широко используется для автоматической генерации текста. Основное применение метода автоматического создания контента – генерация дорвейного бизнеса.

При генерации текста самый простой способ выбора следующей буквы из множества возможных следующих букв – это равновероятный выбор любой буквы. Второй – это подсчет частоты встречаемости букв в языке. Третий – цепи Маркова, где исследуются не отдельные буквы и слова, а их пары. Цепи Маркова показали себя на практике как один из наиболее эффективных способов для решения задачи генерации текста.

Способы генерации цепями Маркова:

Берется определенный текст. Выбирается N последовательных слов из текста, которые будут являться первыми сгенерированными словами случайного текста. Далее каждое последующее слово определяется следующим образом. Берутся последние N сгенерированных слов и рассматриваются все вхождения этой фразы в исходный текст. Рассматриваются все слова, которые идут в этом тексте сразу после этих вхождений, и из них выбирается случайное — оно и будет следующим словом сгенерированного текста.

В качестве текстовой массы можно использовать результаты поиска серверов. В качестве первой фразы – фразу для поиска. В качестве продолжения выбирается случайная фраза из результатов поискового сервера.

Для формирования слов могут использоваться частотные характеристики средней длины слова и длины предложения.

Рассмотрим пример 1 для фразы «Марковские модели для создания псевдо контента веб-сайтов». Частота встречаемости для нее приведена в табл. 1.

Таблица 1

Частота встречаемости для примера 1

Первая буква	Следующая буква (частота встречаемости)					
А	р(1)	й(1)	н(1)	пробел(1)		
О	в (2)	д(1)	з(1)	пробел	н(1)	
И	е(1)	пробел	я(1)			
Е	пробел	л(1)	в(1)	н(1)	б(1)	
Я	пробел(2)					
В	с(1)	д(1)	е(1)			
М	а(1)	о(1)				
Р	к(1)					
К	о (2)	и(1)				
С	к(1)	о(1)	е(1)	а(1)		
Д	е(1)	л(1)	а(1)	о(1)		
Л	и(1)	я(1)				
З	д(1)					
Н	и(1)	т(2)				
П	с(1)					
Т	е(1)	а(1)	о(1)			
Б	дефис					
Й	т(1)					
Пробел	м	д	с	п	к	в
Дефис						

Для буквы «А» существует следующая ей последовательность (р, й, н, пробел). Для буквы «В» – (с, д, е). Для буквы «О» – (в, д, з, пробел, н, в) и так далее. То есть для буквы «О» буква «в» встречается в два раза чаще, чем остальные «д, з, пробел, н» буквы.

Построим марковскую модель, которая в зависимости от контекста буквы может описывать, слова, предложения, абзацы, называемый каждый в отдельности блоком текста.

Матрица вероятностей переходов имеет вид:

$$P=[p_{i,j}]=\begin{array}{c|cccc} & S_1 & \dots & \dots & S_n \\ S_1 & P_{1,1} & \dots & \dots & P_{1,n} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ S_n & P_{n,1} & \dots & \dots & P_{n,n} \end{array},$$

где $p_{i,j}$ – вероятность появления j-го блока текста после i-го.

Значения матрицы могут принимать значения в интервале от нуля до единицы. Значение «ноль» соответствует отсутствию такой комбинации в оригинале.

Множество состояний описывается множеством $S = \{S_1, ..., S_n\}$. Мощность множества соответствует количеству уникальных блоков текстовой массы. Мощность множества соответствует количеству уникальных пар блоков текстовой массы. Граф состояний показан на рис. 1.

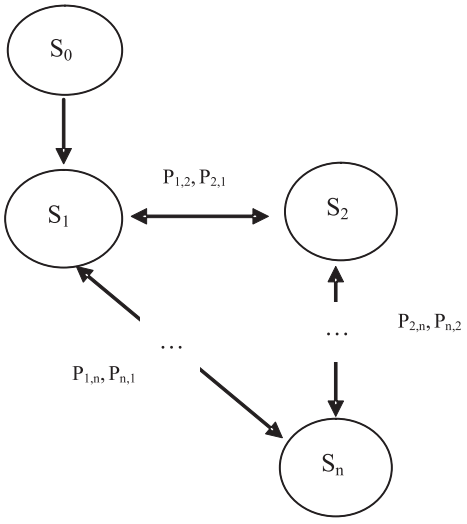

Рис. 1. Граф марковской цепи

При формировании моделей вектор начальных вероятностей выглядит как {1, 0, ..., 0}. Стартовой точкой является начало текстовой массы. При формировании текстовой массы по модели стартовой точкой может являться любое состояние множества S, а именно состояние, содержащее ключевые блоки текста. В случае асинхронного, параллельного процесса начальных состояний может быть несколько.

Марковская модель может моделировать построение слова, словосочетания, предложения, абзаца и целого текста как отдельно, так и в совокупности, представляя собой большой массив данных. С матрицей вероятностей переходов может быть связана матрица аттрибутов, которая содержит различные семантические и синтаксические характеристики блоков текста и их взаимосвязей, их онтологическое представление в реальном мире, то есть база знаний.

Таким образом, метод генерации псевдо – контента заключается в следующем:

1) Построение марковской модели блоков текстовой массы ArrayText.

$\forall$ Connect (X, Y) $\Rightarrow$

$\Rightarrow$ PopulateArrayText (X, Y, Connect (X, Y)),

где X, Y – блоки текстовой массы.

2) Формирование матриц аттрибутов ArrayAttributeRelation (X, Y), ArrayAttributeUnitText (Z),

где Z – блок текстовой массы.

3) Сканирование марковской модели, последовательно идя со стартовой точки, формируя слова, предложения, абзацы до указанной длины.

BuildUnitText (LengthText, ArrayText,

ArrayAttribute) = $\exists$ ArrayText(X, Y) $\wedge$

$\wedge \exists$ ArrayAttribute(X, Y) $\wedge \exists$ LengthText.

4) Обвертывание полученной текстовой массы форматом для отображения в веб.

BuildWebTemplate = AddWebTemplate

(BuildUnitText (LengthText, ArrayText,

ArrayAttribute)).

Для формализации использована логика предикатов первого порядка.

Рассмотрим пример 2. Построим марковскую модель для слова «контент». Множество состояний S = {к, о, н, т, е}. Вектор начальных вероятностей $\pi = \{1, 0,...,0\}$. Граф состояний для примера имеет вид, показанный на рис. 2.

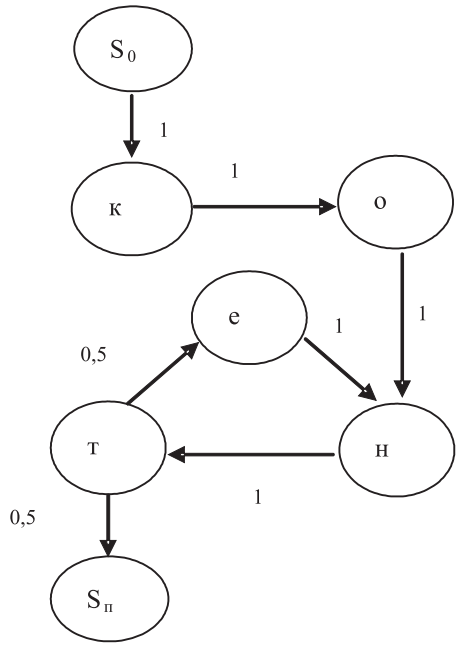

Рис. 2. Граф марковской цепи для примера 2

Матрица вероятностей переходов:

$$P=[p_{i,j}]=\begin{array}{c|ccccccc} & к & о & н & т & е & н & т \\ \hline к & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ о & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ н & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ т & 0 & 0 & 0 & 0 & 0,5 & 0 & 0 \\ е & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ н & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ т & 0 & 0 & 0 & 0 & 0 & 0 & 0,5 \end{array}.$$

Выводы

Таким образом, в результате проведенного исследования проанализированы существующие способы генерации контента, усовершенствована марковская модель, описывающая зависимость блоков текста, предложен метод формирования блоков псевдо текста, использующий предложенную модель.

Полученные результаты применяются для формирования текстов-дорвеев. К перспективе дальнейшего исследования относится исследование семантики сгенерированных текстов.

Список литературы: 1. *Нартова, А.* Прописка в Интернете [Текст] / А. Нартова, А. Набережный // Мир ПК. 2002. – №12. – С. 5-9. **2.** *Керк, Д.* «Вредный» поиск. [Текст] / Д. Керк // ComputerWorld. – 2007. – №15. – С. 37. **3.** *Хан, У.* Системы автоматического реферирования [Текст] / Удо Хан, Индерджиет Мани // Открытые системы. – 2000. – № 12. – С. 18-22. **4.** *Корявченко, А.* Алгоритмы кодогенерации. Кодогенерация при программировании с использованием платформы Microsoft.NET. [Текст] / А. Корявченко // RSDN Magazine. – 2003. – № 4. – С. 16-24. **5.** *Коробова, И.Л.* Автоматизированная система синтеза текста на основе технологии шаблонизации [Электронный ресурс] / ГОУ ВПО "Тамбовский государственный технический университет "Национальный фонд подготовки кадров Педагогический Интернет-клуб Тамбовской области // Материалы межрегион. науч.-практ. конф. «Информатизация системы образования Тамбовского региона». – Режим доступа : http://club-edu.tambov.ru/main/news/index.php?r=konf1&f=t12, свободный. **6.** *Рассел, С.* Искусственный интеллект: современный подход, 2-е изд. [Текст] / С. Рассел, П. Норвиг. – Москва: Изд. дом «Вильямс», 2006. – 1408 с.

Поступила в редколлегию 06.04.2010 г.

УДК 004

Марківські моделі для створення псевдо-контенту вебсайтів / М.В. Збітнєва //Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 1 (72). – С. 104–108.

Генерація контенту знаходить своє призначення у різних галузях. В статті проаналізвано способи генерації контенту, наведено марківські моделі, архітектуру інтеллектуального агенту, а також метод формування вебконтенту.

Табл. 1. Іл. 2. Бібліогр.: 6. найм.

UDK 004

Markov models for web-site content pseudo-generation. / M.V. Zbitneva // Bionics of Intelligence: Sci. Mag. – 2010. – № 1 (72). – P. 104–108.

Content generation is used in different domain. There are analized ways of content generation, presented markov models, agent architecture, and method forming web-content.

Tab. 1. Fig. 2. Ref.: 6. items.