

РЕДАКТОР-НАВИГАТОР ЭЛЕКТРОННЫХ КАРТОТЕК

М.Г. Мальковский, д.ф.-м.н.; С.Ю. Соловьев, д.ф.-м.н. (МГУ им. М.В. Ломоносова),
А.Н. Сотников, д.ф.-м.н. (МСЦ РАН, г. Москва)

Творческий процесс порождения новых идей зачастую скрыт от сознания и изолирован от воли человека. Тем не менее пассивное ожидание вдохновения совершенно бесплодно. Вдохновение оставляет следы только на подготовленной почве в головах у тех, кто терпеливо, но активно ждет его. Общепризнанным занятием на период *ожидания идей* является ведение личной картотеки по теме ожидания [1]. Картотека как систематизированное собрание однотипных бумажных карточек известна с давних времен. На карточках записываются цитаты, тезисы, ссылки, новости и пр. Дальнейшая работа с картотекой состоит в ее постоянном пополнении и изменении, а также в кластеризации и классификации накопленных карточек. А однажды группа карточек сама собой превращается в окончательный продукт интеллектуальной деятельности: в статью, в рецепт, в словарь и т.д.

На первый взгляд для поддержания картотек существует большое количество разнообразных программных продуктов, от почтовых клиентов до организаторов. Однако все эти продукты в большей степени ориентированы на первоначальное порождение картотеки, чем на систематическую деятельность по ее ведению. В связи с этим интерес представляют принципы организации многофункционального редактора-навигатора, предназначенного для творческой работы с электронными карточками и генерации интеллектуальных произведений, завершающих эту работу.

Концепция редактора-навигатора исходит из того, что:

- электронные карточки связаны между собой бинарными отношениями, типы которых известны заранее;
- электронная карточка представляет собой последовательность символьных строк, каждой из которых приписаны цвет переднего плана и цвет фона;
- электронная карточка допускает однозначный разбор на группы одинаково окрашенных, подряд расположенных строк, в дальнейшем именуемых *цветолями*;
- набор цветолей, образующих карточку, и семантика каждого из них известны заранее.

Считается, что каждая карточка, помимо содержательного наполнения, имеет обязательный блок служебной информации, включающий формальный уникальный идентификатор и различные метки, и необязательное (осмысленное) уникальное наименование.

Бинарное отношение, связывающее две карточки, задается посредством записи в одну из кар-

точек уникального наименования другой карточки. Запись производится в цветоле, однозначно соответствующий типу отношения. Фактически отношения моделируются посредством ссылок, а типы отношений определяются стартовыми цветолями. При таком подходе изменения связей между карточками и их содержимого сводятся к редактированию специально устроенного текстового файла (Т-фрагмента) с последующей его интерпретацией.

Классификация функциональных возможностей редактора-навигатора имеет следующий вид: функции копирования карточек в Т-фрагмент (выбора, поиска, проверки); функции модификации Т-фрагмента (интерактивного редактирования и пакетной обработки, в том числе функции анализа); функции (локального) преобразования картотеки (загрузки Т-фрагмента в картотеку и экспорта карточек).

Многие функции редактора-навигатора реализуются в виде самостоятельных утилит, подключаемых посредством технологии *plug-in*. Вообще говоря, редактор-навигатор можно рассматривать как совокупность сведенных под одну крышу утилит, использующих и модернизирующих (по мере необходимости) единый индекс картотеки. Рассмотрим основные функциональные возможности редактора-навигатора.

Функции копирования карточек в Т-фрагмент различаются механизмами выявления карточек, подлежащих копированию.

Функции выбора позволяют пользователю перемещаться по структуре картотеки, заданной бинарными отношениями, и поштучно выбирать карточки. Формируемый редактором-навигатором набор отправных точек обеспечивает доступность любой карточки за конечное число прямых и обратных переходов по ссылкам-отношениям.

Функции поиска позволяют находить карточки по фразам, заданным для тех или иных цветолей. Дополнительно функции поиска могут обнаруживать в картотеке двойников выделенного подмножества картотек. Понятие двойника уточняется в поисковом запросе.

Функции проверки выявляют подозрительные карточки на наличие дефектов. Открытый список проверок, в частности, включает:

- выявление карточек без завершающей точки в цветоле, предназначенном для размещения заметки;
- выявление двух одинаковых подряд расположенных слов;
- выявление связей-дубликатов.

Конструктивный подход к устранению ошибок – важнейшее условие существования больших картотек. Вместе с тем утилиты, реализующие упрощенные алгоритмы проверки, могут раз за разом объявлять подозрительными одни и те же карточки, вполне качественные с точки зрения пользователя. Универсальный механизм защиты проверенных пользователем карточек от повторных проверок использует метки, которые автоматически размещаются в блоках служебной информации проверенных карточек, отменяют последующие проверки и существуют до момента очередного редактирования карточек.

Функции модификации Т-фрагмента реализуются посредством текстового редактора и утилит пакетной обработки Т-фрагмента.

Редактор, ориентированный на работу с карточками, дополнительно к стандартным функциям правки текста предоставляет пользователю некоторые специальные операции:

- создать/удалить цветоль;
- открыть служебную информацию для просмотра/редактирования;
- скрыть служебную информацию;
- удалить карточку;
- породить новую карточку.

Операции пакетной обработки реализуют отдельные преобразования Т-фрагмента:

- стандартное форматирование карточек;
- разделение по определенным правилам больших по объему карточек на несколько малых;
- установление однотипных связей для карточек из Т-фрагмента;
- факторизация множества карточек по заданному цветолу и т.п.

Кроме того, пакетная обработка может использоваться для поддержки творческих операций. Реализованные с этой целью функции не изменяют Т-фрагмент, но позволяют пользователю по-новому оценить его. К таким функциям относятся: эвристическая сортировка карточек, визуализация бинарных отношений, представление электронной картотеки в бумажном виде и др.

Функции преобразования картотеки позволяют загружать (исправленный) Т-фрагмент в картотеку, а также экспортировать карточки, попавшие в Т-фрагмент, из картотеки во внешний файл. Экспорт реализуется как частный случай применения загрузки.

Утилита загрузки исследует Т-фрагмент на корректность и в зависимости от результатов исследования либо возвращает его на доработку, либо исполняет, изменяя содержимое картотеки. Алгоритм загрузки состоит из пяти последовательных этапов.

Этап 1. Формирование списка уникальных идентификаторов для карточек из Т-фрагмента. Если совпала хотя бы одна пара идентификаторов, то загрузка отменяется.

Этап 2. Проверка условия уникальности наименований карточек с учетом предполагаемого изъятия части карточек и включения новых. Если обнаруживается пара карточек с совпадающими наименованиями, то загрузка отменяется.

Этап 3. Проверка корректности бинарных отношений. Если обнаруживается, что в результате загрузки появляется связи в никуда, то загрузка отменяется.

Этап 4. Все карточки с уникальными идентификаторами, выявленными на этапе 1, отмечаются как удаленные.

Этап 5. В картотеку переписываются карточки из Т-фрагмента и при этом изменяются и/или удаляются метки проверенных карточек.

Существуют два обстоятельства, существенно усложняющих реализацию четвертого этапа.

Если реализуется коллективная картотека, то есть предполагается, что картотеку могут одновременно редактировать несколько пользователей, то карточки должны иметь индивидуальные номера версий, изменяемые после каждой модификации, и попытка удалить карточку с номером версии, отличным от номера, скопированного в Т-фрагмент, отменяет загрузку.

Если удаляется карточка, которая, судя по меткам, существует в картотеке достаточно давно и для которой, судя по Т-фрагменту, не предполагается никакой замены, то желательно запросить подтверждение операции. В противном случае можно потерять важную информацию по элементарному недосмотру.

Работа пользователя с редактором-навигатором сводится к повторению (с вариациями) группы из трех операций: формирование → редактирование → загрузка.

Таким образом, шаг за шагом нарабатывается структурированный массив карточек, в котором роль разделов исполняют созданные с этой целью специальные карточки. Важно, что одна и та же карточка может участвовать в нескольких разделах.

Описанный редактор-навигатор является главным звеном технологии подготовки терминологической сети проекта www.glossary.ru – Служба терминологических толковых словарей [2]. По состоянию на начало 2008 года сеть объединяет 52 тысячи карточек с определениями терминов научной и деловой лексики. Конечно, с ростом объемов данных технологию придется менять кардинальным образом, но, по субъективным ощущениям, это событие произойдет не ранее достижения объема в 120 тысяч карточек. Тогда же проявятся и подходы к реализации редактора-навигатора сверхбольших электронных картотек.

Список литературы

1. Альтшуллер Г.С. О личной картотеке // Теория и практика решения изобретательских задач /Под ред. Г.С. Альтшуллера - Горький, 1976. - С. 178-181.

2. Мальковский М.Г., Соловьев С.Ю. Универсальное терминологическое пространство // Тр. междунар. сем. «Компью-

терная лингвистика и интеллектуальные технологии». - М: Наука, 2002. - Т.1. - С. 266-270.

СИСТЕМА ФИЛЬТРАЦИИ ИНТЕРНЕТ-ТРАФИКА НА ОСНОВЕ МЕТОДОВ DATA MINING

(Работа поддерживается грантами: РФФИ, № 06-01-00691; поддержки научных школ, № 02.445.11.7427 и Президента РФ, МК-4264.2007.9)

*В.В. Глазкова; В.А. Масляков; И.В. Машечкин, д.ф.-м.н.; М.И. Петровский, к.ф.-м.н.
(МГУ им. М.В. Ломоносова)*

Проблема контроля доступа к Интернет-ресурсам актуальна и имеет важное прикладное значение по следующим основным причинам: блокирование доступа к нелегальной информации, предотвращение доступа к Интернет-ресурсам в личных целях в учебное или рабочее время, предотвращение утечки конфиденциальной информации через Интернет.

Так как основная часть Интернет-ресурсов сегодня является динамической, это делает малопригодными традиционные сигнатурные подходы, основанные на базе данных статических URL или осуществляющих анализ содержимого на основе ключевых слов. Поддержка подобной базы данных в актуальном состоянии в силу постоянного изменения содержимого Интернет-ресурсов существенно снижает точность работы сигнатурных систем. Поэтому необходим некий подход, который позволил бы быстро и точно в режиме реального времени осуществлять анализ содержимого Интернет-ресурсов.

Авторами статьи предлагается подход к построению систем фильтрации трафика на основе методов машинного обучения, а именно многоклассовой многоотечной классификации гипертекстовых документов. Данный подход имеет ряд существенных преимуществ перед традиционными подходами: независимость от внешних баз данных и экспертов; автоматический процесс обучения; высокая скорость и точность работы; обучаемость и дообучаемость; адаптация к потребностям конкретных организаций.

Архитектурное решение

Предлагаемый авторами подход требует разработки как самой архитектуры системы, основанной на методах машинного обучения, так и специализированных алгоритмов разбора и анализа Интернет-документов в силу их специфики.

Для оценки качества функционирования систем фильтрации Интернет-трафика используются следующие количественные показатели (характеристики):

- *точность анализа* – процент верно отфильтрованных Интернет-ресурсов;
- *излишнее блокирование или ложноположи-*

тельные ошибки – процент хороших ресурсов, ошибочно запрещенных системой фильтрации;

- *недостаточное блокирование или ложнопотрицательные ошибки* – процент плохих ресурсов, ошибочно разрешенных системой фильтрации;

- *скорость анализа* – максимальный объем данных, который система может проанализировать в единицу времени.

Проанализировав характеристики и недостатки подобных систем, авторы предлагают следующие требования к современным системам фильтрации Интернет-трафика:

- анализ и фильтрация Интернет-информации в режиме реального времени с возможностью адаптироваться к постоянной динамике изменения Интернет-ресурсов;
- осуществление фильтрации Интернет-ресурсов с использованием как их содержимого, так и метаданных;
- возможность персонифицировать трафик, то есть максимально точно идентифицировать пользователя, от которого поступил запрос на ресурс;
- анализ не только входящего, но и исходящего от пользователей трафика;
- возможность удаления категорий фильтрации и добавления новых категорий с новыми обучающими примерами;
- независимость от языков Интернет-ресурсов.

Авторами предложена архитектура и алгоритмы анализа и фильтрации Интернет-трафика на основе методов машинного обучения для многоотечной классификации гипертекстовых документов [1,2]. Методы машинного обучения обладают достоинствами самообучаемости, адаптируемости, автономности.

Учитывая требования к системам фильтрации трафика, предложена архитектура системы, состоящая из следующих основных компонент.

Ядро системы, координирующее работу модулей системы и сохранение в базе знаний информации о пользователях, запросах, запрашиваемых ресурсах, а также результатов анализа и классификации.