У статті розглядається рішення задачі підвищення ефективності пошуку шляхом створення моделі самонавчальної системи-помічника, мета якої - пошук інформації, цікавої користувачу, і її відображення у вигляді документа, що складається з адаптивних Web-сторінок

Ключові слова: агент, онтологія, адаптивні Web-сторінки

В данной статье рассматривается решение задачи повышения эффективности поиска путем создания модели обучающейся системы-помощника, цель которой - поиск информации, интересной пользователю, и её отображение в виде документа, состоящего из адаптивных Web-страниц

Ключевые слова: агент, онтология, адаптивные Web-страницы

This article represents the decision of searching problems – the model learning system. It`s main goal is to search information, interesting to the user, and displaying it in the form of a document consisting of adaptive Web-pages

Keywords: agent, ontology, adaptive Web-page

УДК 004.853

МОДЕЛЬ ПОСТРОЕНИЯ АДАПТИВНЫХ WEB-СТРАНИЦ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА СЕТИ INTERNET

О.М. Почанский
Аспирант
Кафедра искусственного интеллекта
Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166
Контактный тел.: (057) 702-13-37
E-mail: pochansky.oleg@yandex.ru

1. Введение

В настоящее время со стремительной быстротой развиваются технологии скоростного доступа в интернете, как проводного (оптоволокно) так и беспроводного (Wi-Fi). В результате, только в Украине на конец первой четверти 2010 г. общее число абонентов широкополосного доступа составило около 2,46 млн, из которых почти 2,2 млн – домашние пользователи (на 210 тыс. больше, чем в IV квартале 2009 г.)[1]. На основании этого можно утверждать, что интернет по популярности сопоставим с телевещанием, а по динамике роста аудитории даже опережает пользователей [2]. Для многих он стал не только средством развлечения, но и основным источником получения различной информации.

В связи с этим, еще в конце 20 века появились первые поисковые системы, которые способны находить нужные данные, по запросу пользователя и отображать результат в виде списка адресов страниц на необходимые ресурсы. Но с развитием интернета, количество информации, которое находится в нем, с каждым днем и даже часом постоянно растет. Это привело к тому, что по запросу пользователя поисковая система выдает огромное количество различных страниц, иногда до нескольких сотен тысяч ссылок за один запрос. В результате этого пользователь тратит огромное количество времени только на поиск нужной ему информации, а ему еще необходимое ее обработать. Не помогает даже наличие внутреннего ранжирования выдаваемых результатов поисковыми системами. При этом, не стоит забывать о вероятности не нахождения или неполноты получаемых результатов, исходя из неточности и некорректности составления запросов пользователями или зашумленности нерелевантными ссылками, которые не содержат искомую информацию. Все это приводит к довольно низкому проценту эффективности получения пользователями необходимых знаний с сети Internet.

В качестве возможного решения указанной проблемы предлагается разработать тематический персонализированный интернет ресурс, состоящий из адаптивных Web-страниц. В этом ресурсе будут находиться различные данные (текст, картинки, видео), соответствующие определенной предметной области. Причем эта информация должна быть максимально актуальна, полезна и интересна для каждого пользователя. Поставка и обработка данных из сети Internet выполняется с помощью специализированной программы-агента (рис. 1).

Иными словами, речь идет о программе-ассистенте, которая выполняет поиск и анализ существующей информации в сети Internet, а затем предоставляет выжимку данных, в виде тематического Web-ресурса, который по форме и содержанию похож на электронную газету. Решение данной задачи и является ключевым моментом, исследуемым в рамках данной статьи.

В следующих разделах рассмотрим ближайшие наработки в данной тематике, а затем ознакомимся с авторским решением.

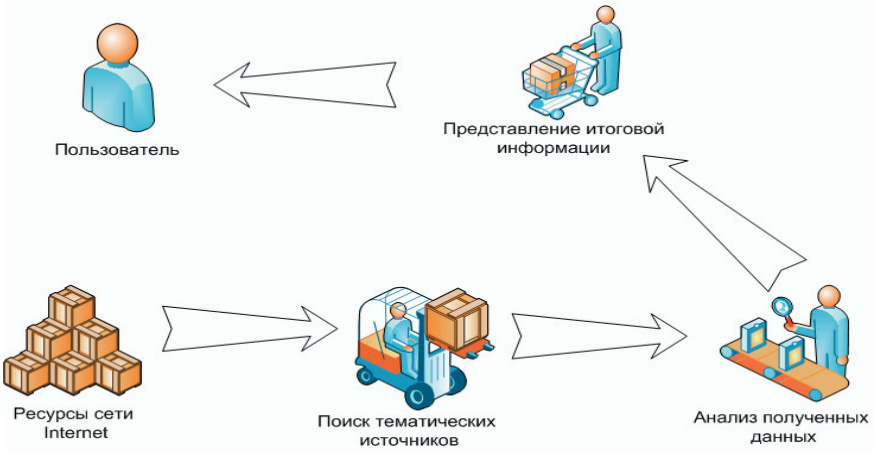

Рис. 1. Общая схема работы программы-агента

2. Анализ существующих моделей

Основа любой модели работы программы-агента зависит от задач, которые он решает. В данном случае основная его задача может быть определена, как поиск различных данных, в соответствии с заранее заданными критериями. А для этого он должен базироваться на эффективных алгоритмах классификации обрабатываемой информации. Рассмотрим их основные особенности.

Традиционные алгоритмы классификации определяют принадлежность одному из классов и схожесть различных документов, основываясь на априорных знаниях о его структуре или используя вероятностные методы, основанные на формулах Байеса [3]. Многие из этих алгоритмов показывают низкую эффективность, когда количество признаков, которые классифицируют документ, существенно больше, чем он сам. В таком случае применяют алгоритмы классификации, которые могут эффективно работать при большом количестве переменных. Они основаны на разделении документов на слова, которые затем записываются в форме графа. По количеству схожих слов, определяется тематическая близость различных документов [4]. Далее остановимся на программах-агентах, построенных на их основе.

Начнем с обзора “NewsAgent”, программы-агента, разработанного в одном из Аргентинских университетов [5]. Данный агент является интеллектуальной программой, которая может генерировать персонализированные газеты (информационные ресурсы), в соответствии с индивидуальными интересами пользователя, основываясь на его анкетных данных и на наблюдениях за ним в режиме реального времени. Для генерирования персонализированных газет он использует статистические методы классификации данных, которые разбивают их на различные классы. При этом каждый класс, в свою очередь, делится на произвольное количество динамических подклассов, используя метод, основанный на сравнениях с эталонными ситуациями. Наблюдение за пользователем выполняется с помощью Java аплета, встраиваемого в браузер, который отслеживает работу пользователя с загруженной в данный момент Web-страницей. Данный агент также фиксирует время, потраченное на прочтение каждой новости из полученной газеты. Таким образом, определяются текущие интересы каждого пользователя, исходя из принципа: чем больше затрачено время на прочтение данного материала – тем более он интересен.

В дальнейшем “новостной агент” был модернизирован [6]. В результате, этого он был раздел на две подпрограммы: “PersonalSearcher” и ”NewsAgent”. Первая подпрограмма отвечает за формирования критериев, по которым определяется интерес пользователя к той или иной тематике. А также за поиск Web документов в соответствии с установленными критериями. Вторая – за генерирование персонализированной газеты на основании данных полученных от первой подпрограммы.

К плюсам данной системы можно отнести:

1. Адаптивность агентов в зависимости от текущих интересов каждого пользователя в режиме реального времени.

2. Автоматическое дробление общих тематик Web документов на более частные. Обеспечивает гибкость процесса определения любимых и не любимых тем в рамках одной предметной области.

3. Максимальное взаимодействие с пользователем с минимальными временными затратами на ответную работу с агентом.

К минусам данной системы можно отнести:

1. Использование статистических методов при классификации Web документов. В результате чего схожесть документов определяется по наличию ключевых слов, а не по их лексическому смыслу.

2. Любой Web документ просматриваемый пользователем, может относиться только к одной общей теме.

3. Заданы ограничения на деление общих тем на более частные в рамках одной предметной области.

4. Не предусмотрено взаимодействие и интеграция с другими системами (к примеру, социальными сетями) с которыми может работать пользователь.

5. Персонализированная газета генерируется в форме списка ссылок и не включает в себе данные из найденных источников.

Далее рассмотри систему InfoStream. Она предназначена для нахождения в сети Интернет новостной информации по интересующим пользователя тематикам, оперативной доставки результатов поиска, предоставления единого интерфейса доступа к информации с тысяч Web-сайтов, минимизации усилий, на отсеивание дублирующейся информации, шума. Даная система обеспечивает интеграцию сетевых информационных ресурсов на базе эффективных средств сбора, обработки, хранения данных и организации эффективного доступа к ним. С помощью неё выполняется автоматизированный сбор информации с Web-сайтов в режиме реального времени, ее структурирование, группировка по семантическим признакам, а также эффективное тематическое избирательное распределение и предоставление доступа к информационным базам данных в поисковых режимах [7]. Результаты работы системы отсылаться на один или несколько электронных адресов пользователя в формате RSS новостей, информационных документов или аналитических отчетов.

К плюсам данной системы можно отнести:

1. Мониторинг основных информационных ресурсов сети Internet в режиме реального времени.

2. Хранение и обновления данных из проиндексированных источников, в формате единой базы данных.

3. Наличие аналитического инструментария – пользователь может в режиме реального времени не только получать результаты поиска, но и формировать дайджесты, строить сюжетные цепочки, анализировать взаимосвязь рубрик, понятий.

К минусам данной системы можно отнести:

1. Отсутствие систем автоматического мониторинга интересов пользователя. Профиль пользователя формируется только в соответствие с заданной им информации и в дальнейшем может корректироваться только им.

2. Дополнительные затраты времени на использование и обучение. Для получения оперативного и эффективного результата, пользователь должен составить расширенный запрос, на специальном языке, который понятен системе.

3. Наличие ограничений на количество тематических запросов пользователя в сутки.

4. Отсутствие возможности объединения полученной информации в формате единого Web-документа.

5. Использования статистических методов при классификации Web документов. В результате чего схожесть документов определяется по наличию ключевых слов, а не по их лексическому смыслу.

На основании проведенного анализа существующих моделей, рассмотрим авторское решение указанного ранее ключевого момента.

3. Постановка задачи и модель ее решения

В качестве решения проблем эффективного поиска данных в сети Internet сформулируем задачу создания универсальной программы-агента или системы, цель которой обеспечить доступ пользователя к интересующей его информации с минимальными потерями времени и максимальной точностью. На основании проведенного анализа существующих разработок, направленных на выполнение указанной выше задачи, было выявлено ряд нерешенных проблем:

1. Реализации методов классификации искомых данных, с учетом их лексических особенностей.

2. Мониторинга интересов пользователя на основании его активности в сети Internet (а не только во время поиска и анализа различных источников информации)

3. Прогнозирования повышения или понижения интереса к той или иной тематике в рамках заданной предметной области.

4. Представления информативного результата в режиме реального времени в формате единого Web-документа.

В качестве возможного варианта решения указанных выше проблем, предлагается рассмотреть модель построения адаптивных Web-страниц на основе интеллектуального анализа сети Internet.

Данная модель основывается на взаимодействии нескольких программ агентов:

1. Web-parser, который отвечает за поиск, анализ и обработку информационных ресурсов сети Internet.

2. Web-monitor, который отвечает за формирование и сопровождение динамического профиля интересов каждого пользователя.

3. Web-constructor, который отвечает за создания адаптивных Web-страниц, в рамках единого персонализированного информационного ресурса с актуальными данными сгенерированными и скорректированными в соответствие с текущими интересами каждого пользователя.

Общая схема возможных взаимодействий данных программ-агентов с пользователем в рамках единой системы рассмотрена ниже (рис. 2). Стрелками показаны области работы каждой программы-агента, а также расписаны возможные действии пользователя.

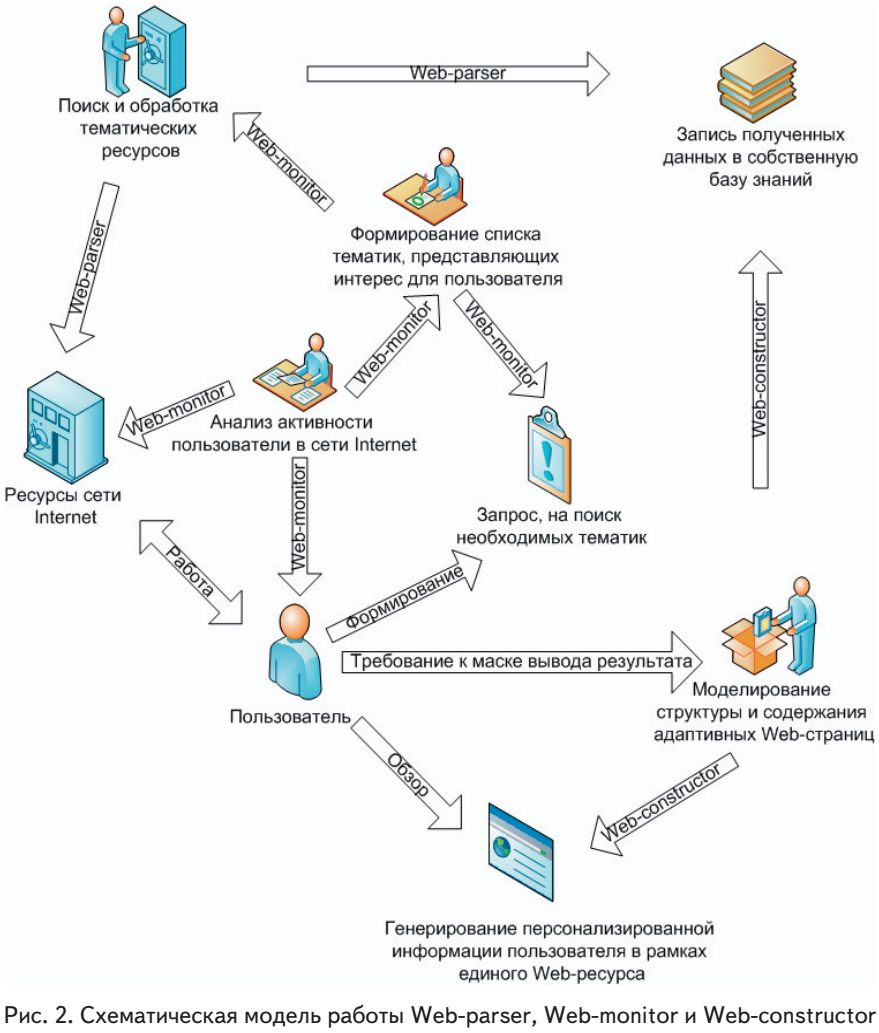

Рис. 2. Схематическая модель работы Web-parser, Web-monitor и Web-constructor

Далее остановимся на некоторых технических особенностях работы указанных выше схемы:

1. Поиск и обработка тематических ресурсов состоит из нескольких этапов. Сначала задается список эталонных Web-ресурсов, из которых будут браться данные для каждого пользователя. Он храниться в заранее созданном rdf-файле в формате: адрес, краткое описание. Затем загружается текущий профиль интересов пользователя, который предварительно был создан Web-monitor. На основании данного профиля выбирается нужные интернет ресурсы, который до этого были преобразованы в онтологию. Структура, онтологии схожа с его навигационной панелью (рис. 3).

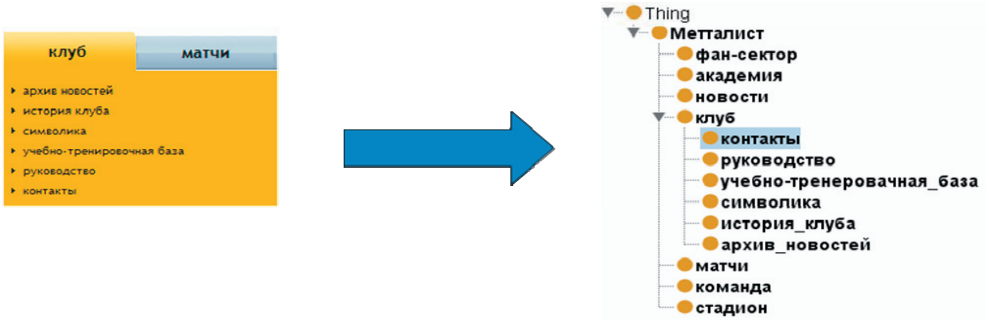

Рис. 3. Представление панели навигации Web-ресурса в виде онтологии с аналогичной структурой

Затем из данной онтологии выбираются данные, которые соответствуют требованиям пользователя.

2. Анализ активности пользователя в сети Internet. Выполняется путем интеграции с различными социальными сетями, в которых зарегистрирован пользователь. Для этого ему необходимо выполнить соответствующие пометки в анкете сгенерированной Web-monitor. Так же данный агент будет определять текущее местоположения пользователя по его ip-адресу.

3. Запись полученных данных в собственную базу знаний. Под базой знаний подразумевается конечное число онтологий сгенерированных из эталонных Web-ресурсов и один rdf-файл, описывающий содержание каждой онтологии.

4. Моделирование структуры и содержания адаптивных Web-страниц. Формируется порядок и формат вывода полученных данных в результате работы Web-parser, с учетом пожеланий пользователя (в форме динамической Web-странички). Данные берутся из базы знаний.

Если все действия были выполнены корректно, то, в конечном счете, пользователь получит разработанный, специально для него, информационный ресурс с интересующей его информацией в заданном им формате. После этого у него появляется возможность составлять статистические отчеты, на основании данных, которые есть на его персонализированном Web-ресурсе.

4. Вывод

В данной статье рассматривается модель решения проблемы низкой эффективности поиска интересующей информации, путем построения адаптивных Web-страниц на основе интеллектуального анализа сети Internet. Основные преимущества данного решения заключается:

1. Отсутствие необходимости задания ключевых слов при классификации документов. Достигается путем преобразования найденных документов в четко определенный формат (онтологию), путем его дробления на информационные единицы (каждый класс онтологии соответствует одной определенной теме).

2. В формировании динамического профиля интересов каждого пользователя на основании его активности в социальных сетях и блогов. Достигается путем добавления информации, оставленной пользователем на этих ресурсах, к его профилю.

3. В отображении результатов в виде единого Web-ресурса в форме подобной с электронной газетой. Достигается путем разбиения результата на структурные блоки (каждый блок описывает определенную тематику). Место размещения каждого из них определяется пожеланиями пользователя или его динамическим профилем интересов.

На данном этапе предложенная модель может быть использована для проектирования и разработки практически систем, направленных на повышение эффективности поиска и создание эталонных Web-ресурсов, не зашумленных ненужной для пользователя информацией.

Литература

1. Домашних пользователей ШПД более 2 млн. // Компьютерное обозрение. – 2010 – №23 – С. 2.
2. Википедия – интернет энциклопедия [http://www.wikipedia.org].
3. Ланде, Д.В. Поиск знаний в Internet. Профессиональная работа.: пер. с англ. – М.: Издательский дом “Вильямс”, 2005. – 272 с.
4. Moore J., Han E., Boley D., Gini M., Gross R., Hastings, K., Karypis, Kumar V. and Mobasher B. Web page Categorization and feature selection using association rule and principal component clustering, 1997.
5. Cordero D., Roldan P., Schiaffino S. and Amandi A. Intelligent agents generating personal newspapers. In Proceedings of the 1st international conference on enterprise information system (ICEIS 99), Setubal, Portugal, 1999. – pp. 195-202.
6. Cordero D., Roldan P., Schiaffino S. and Amandi A. Interface agents personalizing Web-based tasks. Cognitive System Research №5, 2004. – pp. 207-222.
7. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно методическое пособие. / А.Н. Григорьев, Д.В. Ландэ, С.А. Бороденков и др. – К.: ООО “Старт 98”, 2007. – 40 с.