УДК 004.853

О.М. Почанский

Харьковский национальный университет радиоэлектроники, Харьков

ПРИМЕНЕНИЕ СТРУКТУРНЫХ ХАРАКТЕРИСТИК WEB-ДОКУМЕНТОВ ПРИ ОЦЕНИВАНИИ ИХ ПРИВЛЕКАТЕЛЬНОСТИ ДЛЯ КОНЕЧНОГО ПОЛЬЗОВАТЕЛЯ

В данной работе рассмотрен метод повышения эффективности поиска Web-документов. Он заключается в создании внутреннего рейтинга, основанного на оценках меры близости структурных характеристик, заданных пользователем (в числовом или символьном виде) и текущих значений электронного источника информации. Полученные параметры вычисляется с помощью применения теории нечетких множеств и кластерного анализа. Их значимость определяется исходя из заданных пользователем весовых коэффициентов.

Ключевые слова: *персонализированный рейтинг, сравнение Web-документов, поиск, семантическое описание, лингвистическая переменная, кластерный анализ, база знаний.*

Анализ проблемной области

С появлением первых Web-страниц на просторах всемирной сети Internet любой человек получил почти неограниченный доступ к информации по различным тематикам, не покидая дома или места работы. Со временем, для многих людей, электронные документы заменили аналогичные по содержанию печатные издания. Некоторые из них рассматривают свой персональный компьютер или мобильное устройство (ноутбук, смартфон) с постоянным доступом в Internet, как единственный (на их взгляд) актуальный и современный источник получения последних новостей, необходимой им информации и новых знаний, отвергая любые другие доступные варианты.

В итоге это привело к бесконтрольному и бесконечному росту количества Web-страниц, посвященных различным тематикам, большинство из которых, к сожалению, содержат низкокачественные или непроверенные данные или просто дублируют друг друга. Такие источники информации в технической литературе часто называют нерелевантными. Кроме того многие из релевантных электронных документов могут быть часто зашумлены [1]. Это объясняется наличием рекламы или ненужной информации в Web-документе, которая не соответствует общей его тематике. Исходя из этого, чем большое количество этих шумов, тем заметно ниже эффективность восприятия итогового текста конечным пользователем [2].

В связи с этим, в современных условиях, чтобы найти информацию, которая соответствуют всем критериям пользователя, необходимо потратить значительное количество времени на обработку различных источников по интересующей его тематике. К сожалению, не улучшают ситуацию и поисковые системы, которые способны выдавать по одному запросу более десяти тысяч релевантных результатов [3], что также не способствует улучшению эффективности поиска (под эффективностью поиска мы будем понимать нахождение нужной информации за минимально допустимое время). При этом так же стоит учитывать несовершенство составляемых пользователем запросов и не всегда их корректное восприятие различными поисковыми системами.

В связи с этим, несмотря на то, что вероятность наличия в Internet необходимой пользователю информации стремится к 100%, вероятность её нахождения обратно пропорциональна общему количеству всех источников в сети [4].

Решать указанную проблему должны поисковые системы, но получаемый ими в итоге результат, исходя из описанных выше причин, часто нуждается в дальнейшей обработке и оптимизации.

Таким образом, на взгляд автора, проблема повышения эффективности поиска является перспективным и интересным направлением исследования, результаты которого будут способствовать улучшению взаимодействия и скорости работы пользователя с ресурсами сети Internet.

Анализ последних исследований и публикаций

Проблеме повышения эффективности поиска посвящено достаточно много работ [5], которые не перестают появляться с завидной тенденцией и в наше время. Это лишний раз подтверждает, что 100% решение, способное преодолеть указанную проблему, так и не было найдено. В большинстве из проанализированных работ [6-9], полученный результат выглядит в форме некого компромисса, при котором пользователю приходится мериться с некоторыми ограничениями или недостатками используемых им систем.

За способом и принципом работы данные системы могут быть разделены на следующие классы:

© О.М. Почанский

1. Автоматизированные поисковые системы, основанные на статических методах нахождения необходимой информации. Они постоянно самостоятельно индексируют всё Internet-пространство, открытое для общего пользования, переходя по всем ссылка, найденным в процессе просмотра и обработки каждой новой (обновленной) Web-страницы, используя специальные программы-агенты, называемые поисковыми роботами. Затем полученные данные группируются по ключевым словам (часто встречаемым в тексте или выделенные автором Web-ресурса, как те, что несут наибольшую смысловую нагрузку в документе), которые взяты из метаданных страниц, и сохраняются на сервере поисковой системы. Вывод результатов выполняется на основании совпадения слов поискового запроса пользователя с ключевыми словами Web-страницы. Те электронные источники информации, в которых данные совпадения были наиболее многочисленными или полными, получают приоритет при выводе. При этом семантика слов и смысловая нагрузка формируемого запроса никак не учитывается при поиске нужной информации и не отражается на итоговом результате. Таким образом, для повышения эффективности поиска, каждой проанализированной Web-странице присваивается своя оценка (рейтинг), которая влияет на её позицию в списке релевантных результатов, выводимых при формировании пользовательского запроса. Одним из примеров успешной реализации таких поисковых систем является: Яндекс [10] и Google [11], которые для формирования внутреннего рейтинга Web-страницы используют взвешенный индекс цитирования (ИЦ) и PageRank соответственно [12]. Первый – численный показатель цитируемости сайта, определяемый на основании уже известных Яндексом электронных документов в сети Internet, второй – мера “важности” страницы, которая зависит от числа внешних ссылок на данную страницу и от их качества (в свою очередь числа внешних ссылок на ссылающие страницы). Также обе системы позволяют задавать ограничения на поиск (по дате, языку, автору). К типичным недостаткам таких поисковых систем можно отнести слишком длинные списки релевантных результатов с достаточно высоким процентом Web-страниц, которые не удовлетворяют текущему требованию пользователя по причине их зашумленности, дублирования данных или просто несоответствие заданному запросу.

2. Автоматизированные поисковые системы, основанные на семантических методах нахождения информации. В отличие от систем, основанных на статистических методах, используют дополнительные модули или XML подобные файлы (онтологии), с помощью которых выявляются зависимости между различными Web-страницами и их тематиками. Основной принцип работы таких поисковых систем заключается в предварительном формировании базы знаний, в которой описаны логические правила (перечень характеристик позволяющих определить лексический смысл исследуемого объекта). С их помощью система “воспринимает” семантический смысл слова, предложения или текста. Таким образом, итоговый результат формируется не на основе совпадения ключевых слов поискового запроса пользователя с искомой Web-страницей, а исходя из наличия общих смысловых связей (преемственность различных тематик). Примерами таких поисковых систем, являются: Kim-Semantic Annotation Platform [7] и Similarity/Closeness-Based Resource Browser [8]. Первая – больше ориентирована на детальную проработку каждой Web-страницы, путем построения обширных цепочек, состоящих из нескольких электронных источников, описывающих схожие тематические объекты. В результате, это позволяет любому пользователю получить всю информацию, которая находится в базе знаний, по интересующему его объекту (выбор возможен только из слов, выделенных в тексте маркером), из любой просматриваемой им Web-страницы. Второй – больше ориентирован на графическую составляющую и репрезентативность. Даная поисковая система позволяет схематически отображать семантическую близость различных источников, на основании смежности различных тематик, которые определяются путем выявления общих элементов, описанных по средствам онтологического инжиниринга [13]. Их близость рассчитываются с помощью расстояния Евклида [9] или нового показателя NGD (Normalized Google Distance [14]) . Главным преимуществом таких систем является возможность предоставлять пользователю не только те результаты, которые удовлетворяют его в текущий момент, но и те, что дают возможность прогнозирования дальнейших пожеланий. К недостаткам таких систем можно отнести ограничения, накладываемые на “глубину” и “ширину” семантического поиска смежных тематик различных источников, а также требования к наличию заранее сформированных файлов, описывающих исследуемую Web-страницу и чувствительность к лексическим особенностям языка её написания. Проверка на зашумленность и возможность установки ограничений на поиск “ненужных ” источников также не выполняется.

3. Автоматизированные поисковые системы, основанные на комбинированных методах нахождения информации. Данные системы используют основные преимущества статистических и семантических методов, для повышения эффективности поиска. Таким примером может выступать система InfoStream [6], которая помимо предоставления результатов поиска, позволяет формировать дайдже-

сты, строить сюжетные цепочки, анализировать взаимосвязь рубрик, динамику понятий. К главному недостатку таких систем можно отнести медлительность поиска, необходимость дополнительной обработки найденных Web-страниц, низкий уровень проверки зашумленности данных и отсутствие возможности установки ограничений на поиск "ненужных" источников.

Общим недостатком поисковых систем описанных выше классов является:

1. Отсутствие эффективных методов борьбы с шумами внутри Web-страниц.
2. Отсутствие расширенных возможностей по установке расширенных ограничений на поиск "ненужных" источников.
3. Пользователь не может участвовать в формировании внутреннего рейтинга релевантных Web-страниц.

Вариант преодоления первого недостатка поисковых систем был детально изложен в авторской работе [1]. Остальные будут рассмотрены непосредственно в следующих разделах данной статьи.

Постановка задачи

Основной целью написания данной статьи является разработка метода, который будет способствовать повышению эффективности поиска в сети Internet различных электронных источников информации в рамках одной предметной области. Для этого необходимо выполнить следующий ряд условий:

1. Составить максимально полный список характеристик, которые будут оптимально, описывать структуру любого Web-документа.
2. Использовать составленный список структурных характеристик Web-документа для прогнозирования итогового результата поиска нужной пользователю информации.
3. Спроектировать базу знаний, в которой будет храниться RDF-файлы с заполненным списком структурных характеристик каждого искомого Web-документа.
4. Сформировать онтологии, которые будут использоваться для установления семантических связей между Web-документами со схожими тематиками.
5. Создать систему автоматизированного наполнения ссформированных онтологий данными, взятыми из базы знаний.
6. Разработать модель сравнения различных Web-документов, посвященных одной тематике.
7. Оценить меру значимости каждого Web-документа, для текущего пользователя, при формировании итогового результата поиска.

Исходя из этого, основной задачей текущей работы будет успешное выполнение поставленной выше цели. Полное решение данной задачи будет изложено в следующем разделе.

Изложение основного материала научных исследований

В предыдущем разделе разрабатываемый метод повышения эффективности поиска в сети Internet различных электронных источников был разбит на ряд подзадач (условий). Каждая из них имеет определенный алгоритм решения, составленный на основании теоретических и экспериментальных исследований.

Полученные результаты приведены в порядке их выполнения в рамках общего решения поставленной задачи.

За основу разработки данного метода был взят ряд некоторых психологических аспектов. В основном они заключается в исследовании влияния внешних характеристик объекта при его выборе из других схожих за функциональностью и назначением предметов. Из этого следует, что в подавляющем числе случаев, человек делает свой выбор исходя из визуальных предпочтений, сформированных на подсознательном уровне. Зная их, можно заранее спрогнозировать итоговый результат или повлиять на него [15]

Следовательно, если таким объектом считать любой Web-документ в сети Internet, то на основании проведенного сравнительного анализа схожих электронных источников информации (экспериментальным путем), обобщенный список его внешних (структурных) характеристик может быть следующим:

1. Размер.
2. Время и дата создания.
3. Объем текста (в страницах).
4. Зашумленность (процентное соотношение шумов ко всему тексту на странице [1]).
5. Количество картинок в тексте.
6. Количество ссылок на внешние источники.
7. Количество посещений (сервис Alexa [16]).
8. Рейтинг в поисковых системах (сервис Seop.ru [17]).
9. Местонахождение (сервис Speed-Tester.Info [18]).
10. Наличие видео.

Приняв за основу установленный список структурных характеристик, можно сформулировать персонализированный макет (шаблон) наиболее подходящего искомого Web-документа, который будет лежать в основе разрабатываемого метода повышения эффективности поиска. При этом пользователю необходимо только задать соответствующие значения (в принятых единицах измерения) и весовые коэффициенты (степень значимости) для каждого из пунктов списка или выбрать их из предлагаемых вариантов.

С учетом того, что значения весовых коэффициентов определяются в диапазоне:

$$\sum_{i=1}^{n} \omega_i = 1,$$

где n – количество структурных характеристик Web-документа.

Так же предлагаемые варианты значений структурных характеристик персонализированного шаблона могут быть представлены как числовыми, так и символьными значениями. К примеру, желаемый "размер" Web-документа может быть определен пользователем фиксированной величиной с точностью до байта или цифровым диапазоном с четко указанными границами. Но существуют некоторые ограничения. Так, для структурных характеристик "местоположения" можно задавать только фиксированные символьные значения (названии региона или страны), а для "наличия видео" – булевы (0 или 1).

В случае если пользователь не может определиться с числовыми значениями, он может выбрать их из символьного списка (предварительно сформированного в качестве вариантов структурных характеристик шаблона Web-документа), который состоит из слов или словосочетаний естественного или искусственного языка. Каждый элемент, которого, по сути, представляет собой нечеткое подмножество универсального множества, образованного из всех его значений. При этом данный список называется лингвистической переменной. А любое нечеткое подмножество универсального множества характеризуется функцией принадлежности, которая ставит в соответствие каждому элементу число из интервала [0, 1], характеризующее степень принадлежности к данному подмножеству [19].

К примеру, структурная характеристика шаблона Web-документа "размер" будет считаться лингвистической переменной, если она принимает символьные значения: большой, средний и малый – каждое из которых является нечетким подмножеством универсального множества Т(Размер):

$$\text{Т(Размер)} = \text{большой} + \text{средний} + \text{малый},$$

где знак "+" является операцией не сложения, а объединения.

Для присвоения данным подмножествам, числовых эквивалентов необходимо использовать функции принадлежности "обобщенный колокол" и "сигмоидная функция".

Функция принадлежности "обобщенный колокол" была выбрана исходя из того, что она симметрична, отличается простотой записи и является наиболее используемой при описании нечетких множеств [20]. С помощью нее можно описывать нечеткие множества измеримых понятий (размер Web-документа, объем текста и т.д.), с четко определенными интервалами допустимых значений. Она зависит от трех параметров и в аналитической форме записывается следующим образом [20]:

$$f(x,a,b,c) = \frac{1}{1+\left|(x-c)/a\right|^{2b}},$$

где с определяет расположение центра функции принадлежности; а и b – оказывает влияние на форму кривой (данные параметры подбираются опытным путем).

Сигмоидная функция принадлежности ассиметрична. Она позволяет описывать нечеткие множества, в случае, когда, по ряду причин, невозможно определить "верхнюю" границу (максимально допустимое значение) интервала измеримых понятий. А также в тех случаях, когда в "нижних" границах интервала степень принадлежности измеримого понятия к нечеткому множеству максимальна [20]. В аналитической форме сигмоидная функция записывается следующим образом [20]:

$$f(x,a,c) = \frac{1}{1+e^{-a(x-c)}},$$

где а – коэффициент крутизны, а с – координата перегиба функции принадлежности (данные параметры подбираются опытным путем).

Для того чтобы построить функции принадлежности для нечетких подмножеств лингвистической переменной "размер", необходимо выполнить ряд условий:

1. Выбрать одну величину для вычисления размера Web-документа. В данном случае это будут мегабайты.

2. Выполнить нормализацию полученных значений размерности Web-документов. Для этого округлим их с точностью до одной сотой, а затем умножим каждое из них на 10.

3. Установить границы, описывающие нечеткие подмножества измеримых понятий универсально множества Т(Размер) и выполнить нормализацию их величин:

$\text{большой} = [1.5\text{ мб}, \infty)$, после нормализации: $\text{большой} = [15, \infty)$;

$\text{средний} = [0.3\text{ мб}, 2.7\text{ мб}]$, после нормализации: $\text{средний} = [3, 27]$;

$\text{малый} = [0, 1\text{ мб}]$, после нормализации: $\text{малый} = [0, 10]$.

Исходя из этого, нечеткие подмножеств универсального множества Т(Размер) определяются функцией принадлежности вида:

$$\mu\text{"áî ëüø î é"}(x) = \begin{cases} 1, & x \geq 28.5; \\ \dfrac{1}{1+e^{-(x-22)}}, & x \in [15, 28.5]; \end{cases}$$

$$\mu\text{"ñðåäíè é"}(x) = \begin{cases} 0, & x \in [3, 5], [20, 27]; \\ \dfrac{1}{1+\left((x-12)/3\right)^{2*3}}, & x \in [5, 20]; \end{cases}$$

$$\mu\text{"ì àëûé"}(x)=\begin{cases}0,\ x\in[7.5,\ 10];\\ \dfrac{1}{1+e^{2(x-5)}},\ x\in[0,\ 7.5].\end{cases}$$

Полученные представления данных нечетких подмножеств отображены на графике выше, в порядке возрастания значений интервала по горизонтальной оси элементов x (рис. 1).

При этом, исходя из установленных границ измеримых понятий, значения нечетких подмножеств универсального множества

T(Размер) μ"большой"(x) и μ"малый"(x) описываются сигмоидной функцией принадлежности, а μ"средний"(x) – функцией принадлежности “обобщенный колокол”.

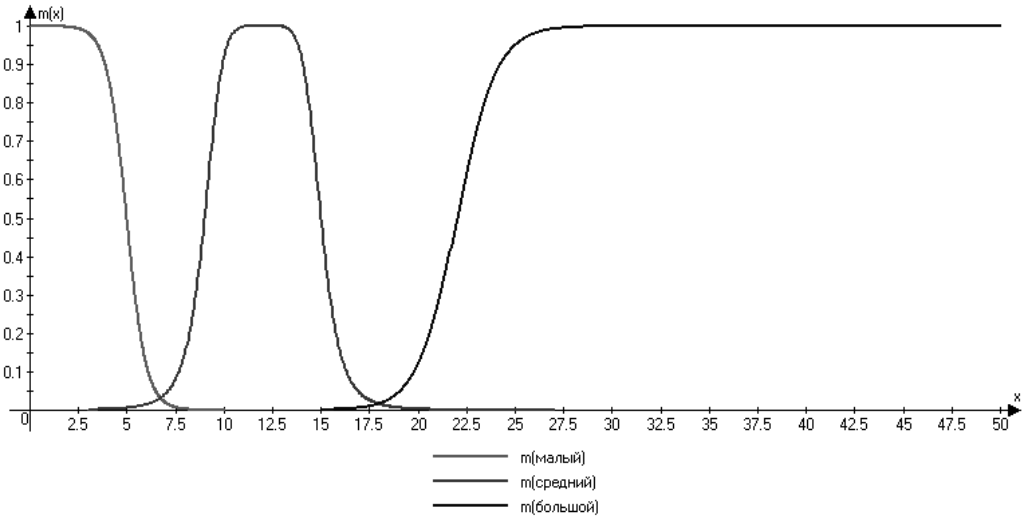

Рис. 1. Функция принадлежности нечетких подмножеств

Затем данные нечеткие подмножества должны быть нормализованы по формуле[21]:

$$\mu(x)=\frac{\mu(x)}{\sup\mu(x)},$$

где $\sup\mu(x)$ – верхняя граница функции принадлежности.

Исходя из графика (рис. 1) $\sup\mu(x)=1$ для всех нечетких подмножеств универсального множества T(Размер). Следовательно, они нормальны [21] и не нуждаются в дальнейших преобразованиях.

Выбор этих функций принадлежностей был сделан опытным путем, на основании использования прямых методов построения. Они характеризуются тем, что эксперт непосредственно задает правила определения значений функции принадлежности $\mu(x)$, описывающей элемент x [22]. В данном случае на начальном этапе роль эксперта выполняет автор данной статьи. В перспективе, для повышения объективности, планируется проводить специализированное анкетирование пользователей, который будут использовать данный метод повышения эффективности поиска Web-документов на практике.

После того, как шаблон структурных характеристик Web-документа был сформирован пользователем, его можно использовать при поиске необходимой ему информации. Следовательно, его необходимо сравнивать с существующими значениями структурных характеристик искомых Web-документов в сети Internet. При этом полученные значения Web-шаблона следует представлять в виде Xml-файла.

Для выполнения данного действия необходимо создать хранилище проиндексированных источников с их Rdf-описаниями, сформированным на основании стандарта Dublin Core [23]. Так же к ним нужно добавлять данные, выявленные при анализе Web-документов [16-18], которые соответствуют элементам списка его структурных характеристик, рассмотренных ранее. В итоге они образуют с Rdf-описаниями общий Xml-файл, характеризующий каждый из исследуемых источников информации. После этого сформированное хранилище может выступать в роли базы знаний основных тематических ресурсов сети Internet. При этом ее наполнение и обновление новыми данными будет выполняться с помощью специализированных программ-агентов, которые детально рассмотрены в статье “Модель построения адаптивных Web-страниц на основании интеллектуального анализа сети Internet” [24]. Так же в данной статье рассмотрен процесс построения семантических связей между Web-документами со схожими тематиками в форме онтологий. Он позволяют структурировать данные в базе знаний (в рамках одной предметной области) не только по ключевым словам, но и по лексическому смыслу.

Таким образом, наиболее подходящий, для пользователя Web-документ, будет, тот, чьи значения структурных характеристик совпадают с заданным персонализированным шаблоном. Но в таком случае полученный список релевантных результатов может быть слишком мал в связи с тем, что вероятность данного события слишком низка из-за огром-

ного количества информационных ресурсов в сети Internet.

Исходя из этого, в качестве итогового результата, предлагается рассматривать все Web-документы, которые есть в базе знаний по искомой пользователем тематике, отсортированные по максимальной близости к заданным значениям персонализированного шаблона, применяя для этого формулу вычисления коэффициента удаленности между различными объектами, называемую Евклидовым расстоянием. Данный выбор был сделан на основании того, что рассматриваемая мера сходимости является самой популярной метрикой в кластерном анализе (для номинальных шкал): она отвечает интуитивным представлениям о близости и, кроме того, очень удачно вписывается своей квадратичной формой в традиционно статистические конструкции [25] В общем виде Евклидово расстояние выглядит как:

$$d_{E_{ij}} = \left(\sum_{l=1}^{n} \left(x_i^l - x_j^l \right)^2 \right)^{1/2},$$

где n – общее число всех признаков объектов; x_i^l – значение l-го признака у i-го объекта; x_j^l – значение l-го признака у j-го объекта; $l = \overline{1,n}$; i, $j = \overline{1,m}$ (m – общее количество исследуемых объектов).

С учетом весовых коэффициентов (ω_n) заданных пользователем для значений структурных характеристик персонализированного шаблона Web-документа Евклидово расстояние примет вид:

$$d_{E_{ij}} = \left(\sum_{l=1}^{n} \omega_n \left(x_i^l - x_j^l \right)^2 \right)^{1/2}.$$

Следовательно, если в качестве примера рассчитать расстояние между Web-документом (Doc) и персонализированным шаблоном (Tem). При условии, что для них определены два произвольных параметра (p1, p2) из списка структурных характеристик. То, в таком случае, можно определить степень их близости в виде числового коэффициента ($d_e(\text{Tem}, \text{Doc})$). В графическом виде это выглядит следующим образом (рис.2).

Из данного графика (рис. 2) видно, что, в данном случае, расстояние близости между Web-документом (Doc) и персонализированным шаблоном (Tem) представлено в виде отрезка $d_e(\text{Tem}, \text{Doc})$ с координатами точек, которые соответствуют значениям параметров (p1 и p2), являющихся структурными характеристиками этих объектов. А его длина определяется по формуле Пифагора для прямоугольного треугольника:

$$d_e(\text{Tem}, \text{Doc}) =$$

$$= \sqrt{(d_{p2}(\text{Tem}) - d_{p2}(\text{Doc}))^2 + (d_{p1}(\text{Tem}) - d_{p1}(\text{Doc}))^2};$$

где $d_e(\text{Tem}, \text{Doc})$ – гипотенуза, а катеты $(d_{p2}(\text{Tem}) - d_{p2}(\text{Doc}))$ и $(d_{p1}(\text{Tem}) - d_{p1}(\text{Doc}))$ проекции данной гипотенузы на каждую из осей графика, которые, по сути, являются расстояниями между значениями объектов при наличии только одного из двух рассматриваемых параметров.

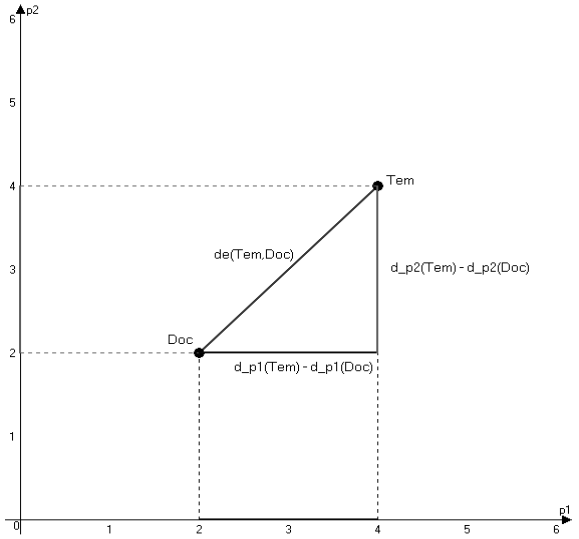

Рис. 2. Графическое представление Евклидова расстояния для двух объектов

Добавление другого Web-документа из базы знаний, соответствующего искомой пользователем тематике с иными значениями параметров p1 и p2, по аналогии, может рассматриваться как новая точка на графике. Затем через нее можно построить отрезок $d_e(\text{Tem}, \text{Doc2})$ (рис. 3), который будет определять меру близости персонализированного шаблона Tem к новому объекту Doc2.

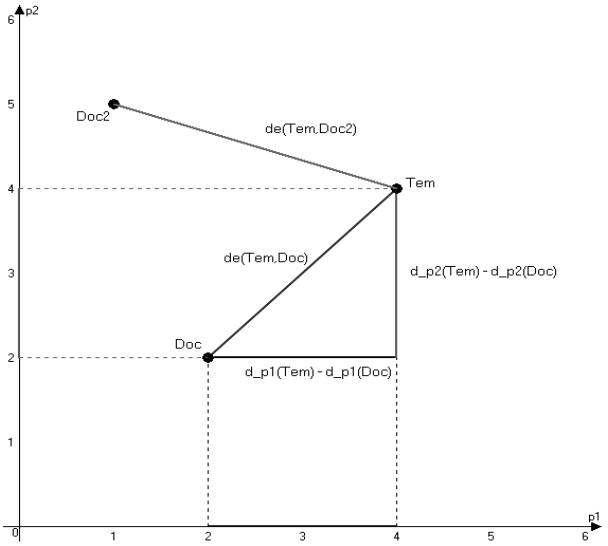

Рис. 3. Добавление нового Web-документа

А постепенное увеличение анализируемых параметров различия объектов приведет к росту осей на графике, на которые будут проектироваться по-

строенный отрезок. Тем самым увеличивается точность определения значений меры близости межу электронным источником и сформированным шаблоном Tem. Использование весовых коэффициентов позволит определить меру значимость каждого из рассматриваемых параметров.

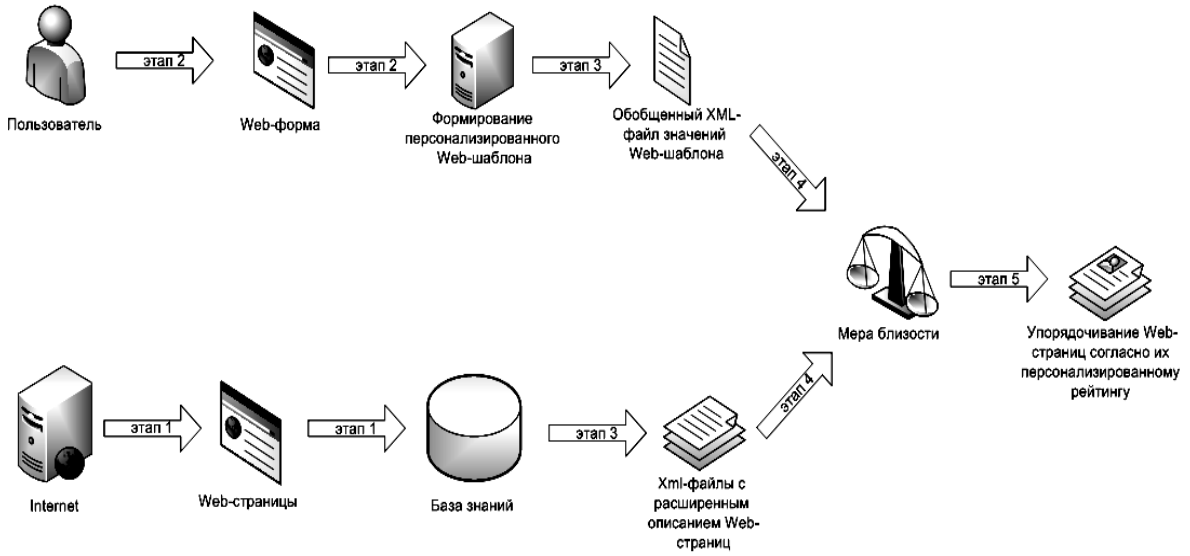

Рис. 4. Обобщенная схема метода повышения эффективности поиска

Исходя из этого, последовательное вычисление расстояния Евклида для всех релевантных результатов приведет к созданию внутреннего персонализированного рейтинга Web-документов, который наиболее соответствуют составленному списку всех значений структурных характеристик. Следовательно, его применение станет более понятным и осмысленным, на взгляд конечного пользователя. В результате это позитивно отразиться на эффективности поиска информации в рамках заданной предметной области.

Таким образом, полученная модель построения персонализированного рейтинга Web-документов является основной функциональной частью в рамках разрабатываемого метода повышения эффективности поиска.

В заключении рассмотрим обобщенную схему работы описанного выше метода (рис 4). Она состоит из нескольких последовательных этапов:

1. Заполнение базы знаний Web-страницам из сети Internet, формирование семантических связей в рамках схожих предметных тематик в форме онтологий.

2. Определение пользователем весовых коэффициентов и задание значений структурных характеристик персонализированного шаблона с помощью заполнения соответствующих полей стандартной Web-формы.

3. Представление персонализированного шаблона и структурных характеристик Web-страниц в виде Xml-файлов.

4. Выявление значений меры близости между Web-страницами и персонализированным шаблоном с помощью формулы Евклидового расстояния.

5. Формирование рейтинга Web-страниц на основании их удаленности от значений шаблона.

Выполнение указанных выше этапов позволит создавать поисковые программы на основе предлагаемого метода. Особенно он будет полезен в разработке автоматизированных поисковых систем, основанных на комбинированных методах нахождения информации.

Выводы

В данной работе рассмотрен метод повышения эффективности поиска Web-документов в сети Internet. Он заключается в формировании персонализированного рейтинга электронных источников информации. Это осуществляется путем описания желаемых структурных характеристик Web-страниц в виде специализированного шаблона, значения которого могут задаваться пользователем в символьном и числовом виде. При этом в случае выбора символьных значений, предусмотрено их автоматическое преобразование в числовые эквиваленты с помощью применения математического аппарата, основанного на нечетких множествах и лингвистических переменных. После этого сформированный шаблон сравнивается со структурными характеристиками Web-страниц, путем использования основных метрик кластерного анализа. А применение весовых коэффициентов позволяет задавать значимость каждого из параметров электронных источников информации. Таким образом, это позволяет добиваться более взвешенных оценок при вычислении меры близости между исследуемыми объектами, исходя из текущих пожеланий пользователя. В результате полученные значения и являются осново-

полагающими параметрами при формировании рейтинга каждого из Web-документов в рамках заданной предметной области. В итоге, время нахождения нужных данных сократиться, а эффективность их поиска повысится.

В перспективе планируется практическая реализация рассмотренного метода повышения эффективности поиска, в рамках специализированной экспертной системы с расширенными возможностями, которые будут заключаться в интеллектуальном анализе и обработке необходимой и значимой информации для конечного пользователя.

Список литературы

1. Почанский О.М. Критерий «чистоты» как основная характеристика модели улучшения работы поискового агента / О.М. Почанский // Математическое и программное обеспечение интеллектуальных систем MPZIS-2010. – Д.: Днепропетровск, 10-12 ноября 2010 г. С. 180-181.

2. Гусев В.С. Google:эффективный поиск. Краткое руководство / В.С. Гусев. – М.: Вильямс, 2006. – 240 с.

3. Ланде Д.В. Поиск знаний в Internet. Профессиональная работа / Д.В. Ланде. – М.: Вильямс, 2005. – 272 с.

4. Холмогоров В. Поиск в интернете и сервисы Яндекс / В. Холмогоров. – СПб.: Питер, 2006. – 123 с.

5. Chia-Hui Ch. A survey of web information extraction / Ch. Chia-Hui, K. Mohammed, R.G. Moheb and oth. // IEEE Transactions on knowledge and data engineering – NJ, USA, 2006. – № 18/10. – P. 1411-1428.

6. Ланде Д.В. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно методическое пособие / Д.В. Ланде, А.Н. Григорьев, С.А. Бороденков и др. – К.: ООО "Старт 98", 2007. – 40 с.

7. Popov B. KIM. Semantic annotation platform / B. Popov, A. Kiryakov, D. Manov // Natural language engineering – Cambridge, 2004. – № 10/3-4. – P. 375-392.

8. Krhriyenko O. Similarity/closeness-based resource browser / O. Krhriyenko, V. Terziyan // Visualization, imaging, and image processing. – Cambridge, July 13 – 15, 2009. – P. 184-191.

9. Naumenko A. Service matching in agent system / A. Naumenko, S. Nikitin, V. Terziyan // Applied intelligent – MA, USA, 2006. – № 25/2. – P. 223-237.

10. Поисковая система Google [Электронный ресурс]. – Режим доступа: http://google.com (дата обр. 18.03.11).

11. Поисковая система Яндекс [Электронный ресурс]. – Режим доступа: http://yandex.ua (дата обр. 18.03.11).

12. Яковлев А.А. Раскрутка и продвижения сайтов: основы, секреты трюки / А.А. Яковлев. – СПб.: БХВ-Петербург, 2007. – 336 с.

13. Thuraisingham Bh. Building trustworthy Semantic Web / Bh. Thuraisingham. – NY,USA: Auerbach Publication, 2008. – 402 p.

14. Cilibrasi R. The google similarity distance / R. Cilibrasi, P. Vitanyi // IEEE Transactions on knowledge and data engineering – NJ, USA, 2007. – №19/3. – P. 370-383.

15. Брунер Дж. Психология познания. За пределами непосредственной информации / Дж. Брунер; пер. с англ. – М.: Прогресс, 1977. – 413 с.

16. Alexa the web information company [Электронный ресурс]. URL: http://alexa.com (дата обращения 18.03.11).

17. Seop.ru search engine [Электронный ресурс]. – Режим доступа: http://seop.ru (дата обращения 18.03.11).

18. Speed-tester.info [Электронный ресурс]. Режим доступа: http://speed-tester.info (дата обращения 18.03.11).

19. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. Заде. – М.: Мир, 1976 – 167 с.

20. Ярушкина Н.Г. Нечеткие интеллектуальные системы в среде SciLab: Методическое указание к лабораторным работам / Н.Г. Ярушкина, Н.Н. Ястребова, А.В. Чекина. – Ул-ск.: УлГТУ, 2009. – 28 с.

21. Поспелова Д.А. Нечеткие множества в моделях управления и искусственного интеллекта / Д.А. Поспелова. – М.: Наука, 1986. – 312 с.

22. Захаров В.А. О выборе методов построения функций принадлежности для формализации задач принятия решений [Электронный ресурс] / В.А. Захаров // Математическая морфология. 2006 . – Режим доступа к статье: http://smolensk.ru/user/sgma/MMORPH/N-12-html/borisov/zakharov/zakharov.htm (дата обращения 18.03.11).

23. Dublin Core Metadata Initiative. [Электронный ресурс]. – Режим доступа: http://dublincore.org (дата обращения 18.03.11).

24. Почанский О.М. Модель построения адаптивных Web-страниц на основе интеллектуального анализа сети Internet / О.М. Почанский // Восточно-европейский журнал передовых технологий – 2010. – № 4/7(46). – С. 66-69.

25. Мандель И.Д. Кластерный анализ / И.Д. Мандель. – М: Финансы и статистика, 1988. -176 с.

Поступила в редколлегию 30.03.2011

Рецензент: д-р техн. наук, проф. С.Н. Герасин, Харьковский национальный университет внутренних дел, Харьков.

ВИКОРИСТАННЯ СТРУКТУРНИХ ХАРАКТЕРИСТИК WEB-ДОКУМЕНТІВ ПРИ ОЦІНЮВАННІ ЇХ ПРИВАБЛИВОСТІ ДЛЯ КІНЦЕВОГО

О.М. Почанський

Розглядається метод підвищення ефективності пошуку, шляхом створення внутрішнього рейтингу, заснованого на оцінках міри близькості структурних характеристик, заданих користувачем (в числовому або символьному вигляді) і поточних значень електронного джерела інформації.

Ключові слова: *рейтинг, що персоналізується, порівняння Web-документів, пошук, семантичний опис, лінгвістична змінна, кластерний аналіз, база знань.*

APPLICATION OF THE STRUCTURAL CHARACTERISTICS OF WEB-DOCUMENTS WHEN EVALUATING THEIR ATTRACTIVENESS FOR THE END USER

O.M. Pochanskiy

We consider a method for increasing search efficiency by creating internal rating based on estimates of the closeness measure of the structural characteristics defined by the user (in numerical or symbolic form) and the current values of electronic information sources.

Keywords: *personalized rating, comparison of Web-documents, search, semantic description, linguistic variable, cluster analysis, base of knowledge's.*