ПРЕДМЕТНО-ОРИЕНТИРОВАННЫЙ ПОИСК ИНФОРМАЦИИ В ИНТЕРНЕТ-РЕСУРСАХ НА ОСНОВЕ МЕТОДА ВЗВЕШЕННЫХ ДЕСКРИПТОРОВ

УДК 519.21.681.142

СТЕНИН Александр Африканович
д.т.н., профессор кафедры технической кибернетики
Национального технического университета Украины
«Киевский политехнический институт имени Игоря Сикорского»
Научные интересы: ситуационное моделирование и оптимизация сложных динамических систем, автоматизированные обучающие системы, идентификация и оптимизация систем.
e-mail: alexander.stenin@yandex.ru

ПАСЬКО Виктор Петрович
к.т.н., доцент кафедры технической кибернетики
Национального технического университета Украины «Киевский политехнический институт имени Игоря Сикорского».
Научные интересы: информационные технологии, идентификация и оптимизация систем.
e-mail: vppasko@ukr.net

ЛЕМЕШКО Вячеслав Анатолиевич
аспирант кафедры технической кибернетики
Национального технического университета Украины
«Киевский политехнический институт имени Игоря Сикорского».
Научные интересы: объектно-ориентированное программирование, информационные системы и информационные технологии, архитектура веб-приложений.
e-mail: slavalemeshko@gmail.com

ВВЕДЕНИЕ

Для успешного выполнения ИТ-проекта недостаточно выбрать эффективные технологии и средства разработки, обеспечить необходимый бюджет и найти квалифицированных разработчиков. В любой организации существуют правила и методики, по которым участники проекта (заказчики, аналитики, разработчики, тестеры, технические писатели) распределяют между собой задачи, взаимодействуют друг с другом, создают проектные артефакты (спецификации, исходный код, документацию). Эти правила могут быть четко организованными или хаотичными, быть формально документированными или существовать в головах проектной команды, но в любом случае именно их совокупность называется процессом разработки.

Известно [1], что процесс разработки инновационных информационных технологий (ИИТ) является многоальтернативным, т.е. возникает проблема многокритериальности, которая, как правило, требует привлечения интеллектуальных систем поддержки принятия решений (ИСППР). Это обусловлено тем, что, во-первых, наличие человеческого фактора в процессе разработки ИИТ вносит большую долю неопределённости, и, во-вторых необходимо посмотреть весь спектр допустимых решений в области использования ИИТ, что требует разработки автоматизированных методов извлечения знаний данной предметной области из различных источников информации, в том числе и из Интернет ресурсов.

ХАРАКТЕРИСТИКА ИНТЕРНЕТ КАК ХРАНИЛИЩА ДАННЫХ.

Интернет в отличие от традиционных информационно-поисковых систем (ИПС) имеет следующие особенности [2, 3].

1. Развитие Интернет как информационного хранилища происходило без учёта потребности поиска документа. Как результат, в Интернет, в отличие от традиционных ИПС, где система хранения документов ориентирована на активный поиск [4], система хранения документов является заданной априори относительно задачи информационного поиска.

2. Интернет представляет собой децентрализованное хранилище документов, не имеющее единого управления организацией и развитием. Сеть Интернет гетерогенна, так как используется не только различные платформы, но и различные стандарты представления информации. Интернет объединяет как современное, так и унаследованные системы. Часть информации храниться в виде, отличном от текста (мультимедиа).

3. Социальная гетерогенность – это 83% коммерческой информации и 6% – научно-образовательной [5]. Кроме того, большой социальный разброс по авторам, аудитории, читателям.

4. Интернет – распределённое хранилище, где время доступа к различным его частям неодинаково и может существенно превосходить время доступа к локальному документу.

5. Объём документов в Интернет оставляет несколько биллионов и превышает объёмы самых больших ИПС и постоянно увеличивается [6]. Большая часть информации, хранимой в Интернете, содержится в базах данных (эта часть называется DeepWeb [7]) и недоступно для большинства существующих промышленных ИПС. По оценкам экспертов количество документов, хранящихся в базах данных Интернет, превышают количество документов хранящихся в промышленных ИПС приблизительно в 500 раз.

АНАЛИЗ ПРОБЛЕМЫ ИЗВЛЕЧЕНИЯ ПРЕДМЕТНО-ОРИЕНТИРОВАННОЙ ИНФОРМАЦИИ ИЗ ИНТЕРНЕТ РЕСУРСОВ

По мере распространения поставщиков услуг доступа в Интернет и снижения цен на такие услуги, всё больше людей получали возможность не только использовать информацию из Сети, но и добавлять её, и, довольно часто, бессистемно. В наши дни, благодаря своей доступности WWW становится одним из популярнейших источников информации.

Наряду с появлением в WWW новой информации, имеет место изменение содержания существующих информационных ресурсов. Страницы, посвященные новостям и другой быстро устаревающей информации, обновляются регулярно. Ввиду децентрализованной структуры WWW, ее архитектура и принципы построения не предполагали какой-либо контроль и упорядочение содержимого. Полезность информации варьируется в весьма широких пределах, так как каждый пользователь имеет возможность размещать любые данные, будь то авторская научная статья или личная информация сомнительного содержания. Если человек хочет получить какую-либо информацию, говорят, что у него возникла информационная потребность. Для её выражения он излагает неформальное описание требуемой информации, например, в виде вопроса на естественном языке. Уже на этом этапе проблематично обеспечить полноту такого описания: в каком контексте интерпретировать фразу, насколько подробный требуется ответ.

Разнообразная слабоструктурированная информация в сети Интернет не может быть успешно использована на практике без эффективного доступа. Так, по оценкам экспертов, около 79% журналистов обращаются к Интернету в поисках новостей, и лишь 20% из них находят ту информацию, которая им необходима. Несмотря на выполняемые большие работы по совершенствованию методов информационного поиска проблему вряд ли можно решить традиционно используемыми методами, такими как, булева модель, векторная модель, интерактивная модель и др. модели поиска. И даже, если представить такую ситуацию, что все существующие проблемы информационного поиска в их традиционной постановке будут решены, то большинство пользователей и, в первую очередь разработчики ИИТ, по-прежнему останутся недовольны, так как извлекаемая ими информация будет релевантной поисковому запросу, а не насущным проблемам.

Для преодоления такого противоречия консорциумом W3C развивается направление Semantic Web. По замыслу его создателей реализация этой парадигмы в

Интернете позволит ИС в какой-то степени понимать содержание информации и выступать «интеллектуальными посредниками (агентами)», способными самостоятельно манипулировать ею по заданию человека.

В настоящее время, в смысле автоматизации информационного поиска, активно ведётся работа по разработке алгоритмов, которые автоматически генерируют программы-посредники. Задача извлечения является сложной, поскольку требуется извлечь не только вид схемы данных, но также и связанную с ним семантическую информацию [8]. Актуальное исследование в области работы со слабо структурированной информацией на основе «интеллектуальных агентов» привели к появлению большого количества альтернативных инструментов их создания. Основной задачей извлечения данных из Интернет-ресурса является получение определённых фрагментов информации из указанных HTML-документов [9, 10]. Эта задача близка к задаче автоматической кластеризации и состоит в поиске разложения HTML-документов $D\{d_1 ... d_n\}$ на классы $C_1 ... C_k$, которые содержат документы со схожей структурой, или, другими словами, в определении базиса признаков $\{e_i\}$, формирующих многомерное пространство, и метода разложения документа по этому базису, то есть вычисление координат $\{w_i\}$. В частности, авторами работы [10] предлагается использовать подход, связанный с вычислением весов термов информационно-поисковых систем и использующих векторную модель представления документов. При этом

$$w_i = f_i / \log(N / k_i), \qquad (1)$$

где f_i – частота встречаемости i-го признака, k_i – количество документов, в которых он встречается, а/ N – общее количество рассматриваемых документов. Для оценки качества кластеризации вводится энтропийная мера. Следует отметить, что такой подход, определяющий значимость терма лишь по частоте, не гарантирует значимость документа по смыслу.

Отсюда, задача разработки методов извлечения информации на основе семантического анализа документов на естественном языке является весьма актуальной.

ПОСТАНОВКА ЗАДАЧИ

Семантический подход является в настоящее время одним из основных путей совершенствования информационно-поисковых систем, т.к. прямое лексическое сравнение запросов с индексами документов полностью не удовлетворяет разработчика. Это объясняется тем, что, как правило, найденные документы обладают либо полисемией (т.е. много лишних слов), либо синонимией (т.е. не все значащие слова извлекаются). Поэтому в рамках семантического подхода предлагается метод взвешенных дескрипторов, позволяющий извлекать наиболее значимые по смыслу и значению документы, весьма близкие к предметной области.

Метод построен на основе идеи базисов Грёбнера [11], в качестве которых используются статистически построенные концептуальные дескрипторы. Данный метод предполагает, что концептуальные дескрипторы в предложениях имеют низ лежащий, «латентный» смысл, который затеняется использованием разных слов. Идеалом при определении базисов Грёбнера будем считать техническое задание на инновационное развитие информационных технологий в конкретной предметной области. Для получения значимых концептуальных дескрипторов воспользуемся законами Джорджа Зипфа, известного американского математика и лингвиста [12].

Согласно этим законам создаваемые человеком тексты построены по единым правилам. Зипф предположил, что природная лень человеческая ведёт к тому, что слова с большим количеством букв встречаются реже коротких слов. На основании этого Зипф вывел два закона:

1. Произведение вероятности обнаружения слова в тексте на ранг частоты является постоянной величиной (C). Причем, значение константы в разных языках различно, но внутри одной языковой группы остаётся неизменным, какой бы текст не был. Так, для русских текстов константа Зипфа $C \approx 0.06 - 0.07$.

2. Для конкретного языка форма кривой Зипфа, связывающая количество слов и их частоту в тексте, является неизменной для любых текстов (рис.1).

Законы Зипфа универсальны. Им, в частности, отвечают характеристики популярности узлов сети Интернет. Объяснение законов Зипфа основано на коррелят-

ционных свойствах аддитивных Марковских цепей со ступенчатой функцией памяти [12].

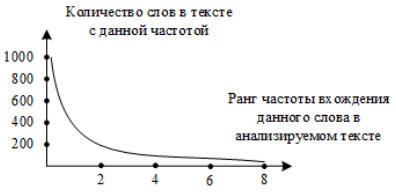

Рис.1. Кривая Зипфа

Анализ второго закона Зипфа показывает, что наиболее значимые слова, а, следовательно, и дескрипторы лежат в средней части кривой Зипфа [13]. Это объясняется тем, что, например, в русском языке наиболее часто встречаются предлоги, местоимения и др., а в английском – артикли и другие. Им отвечает левая часть диаграммы. Правая часть диаграммы соответствует дескрипторам, не имеющим решающего смыслового значения. Следовательно, от того как будет определён диапазон наиболее значимых дескрипторов, зависит успешность работы поисковой системы.

МЕТОД ВЗВЕШЕННЫХ ДЕСКРИПТОРОВ

Во-многом, определение диапазона наиболее значимых дескрипторов зависит от корректного составления двух специальных словарей – тезауруса и стоп-словаря. Тезаурус данной предметной области даёт возможность корректно определить набор концептуальных дескрипторов, т.е. набор наиболее значимых смысловых понятий данной предметной области. Стоп-словарь отсекает «помехи» в виде «лишних» слов, т.е. для русского языка – это частицы, предлоги, местоимения и др.

Количество концептуальных дескрипторов определяется частотным диапазоном выбранных из тезауруса данной предметной области дескрипторов и скорректированных техническим заданием на разработку ИИТ. Как правило, частотный диапазон определяется анализом частоты их появления в техническом задании. Таким образом, пусть мы сконструировали и отобрали n дескрипторов. Тогда по их запросу в Интернет мы получим прямоугольную матрицу

«дескрипторы-документы» $A = \{a_{ij}\}$ размерностью $N \times n$, где a_{ij} – частота появления

i-го дескриптора в j-ом документе, $i = \overline{1,n}$, $j = \overline{1,N}$.

Так как количество документов может оказаться весьма велико, то предлагается провести k-аппроксимацию на основе латентно-семантического анализа (ЛСА).

Латентно-семантический анализ – это метод обработки информации на естественном языке анализирующий взаимосвязь между набором документов и встречающимися в них терминами. В основе метода ЛСА лежат принципы факторного анализа, в частности, выявление латентных связей изучаемых явлений или объектов [14]. При кластеризации документов ЛСА использует для извлечения контекстно-зависимых значений лексических единиц при помощи статистической обработки большого объёма текстов.

В этом смысле, ЛСА можно сравнить с простым видом нейронной сети, состоящей из трёх слоёв: первый слой содержит множество слов, второй– множество документов, соответствующих определённым ситуациям, а третий, средний, скрытый слой, представляет собой множество узлов с различными весовыми коэффициентами, связывающими первый и второй слои.

Для обработки полученных из Интернет документов с целью отбора наиболее значимых в смысловом содержании воспользуемся k – аппроксимацией. Основная идея k – аппроксимации латентно-семантического подхода состоит в замене матрицы A некоторой матрицей $\tilde{A}$, содержащей только k - первых линейно-независимых компонент матрицы A, и отражающей основную структуру различных зависимостей, присутствующих в A.

Говоря более формально, согласно теореме о сингулярном разложении [14] матрица A может быть разложена на произведение трёх матриц:

$$A = USV^T, \qquad (2)$$

где: матрицы U и V - ортогональны, а S – диагональная матрица, значение на диагонали которой представляют собой сингулярные значения матрицы A.

Такое разложение обладает замечательной особенностью: если в матрице S оставить только k наибольших сингулярных значений, а в матрице U и V – только соответствующие этим значениям столбцы, то произведение получившихся матриц S, U и V будет наилучшим приближением исходной матрицы к матрице $\tilde{A}$ ранга k.

$$\tilde{\mathrm{A}} \approx \mathrm{A} = USV^T. \quad (3)$$

Как правило, выборка размерности k зависит от поставленной задачи и выбирается на основе различных статистических критериев, в частности, критерия релевантности следующего вида:

$$R_j = \sum_{i=1}^{n} \alpha_i w_{ij},\ j = \overline{1, N}, \quad (4)$$

где: w_{ij} - частотный коэффициент значимости, α_i – смысловой коэффициент значимости.

В качестве частотного коэффициента значимости можем использовать общепринятую формулу (1), в которой под f_i будем понимать частоту появления i-го дескриптора в j-ом документе. Смысловой коэффициент w_{ij} значимости можно определить либо на основе экспертных оценок смысловой значимости i-го дескриптора в техническом задании, либо как вероятность употребления в техническом задании дескриптора t в данном смысловом значении m_i, вычисляемую по формуле

$$P_t(m_i) = \frac{c(t, m_i)}{\sum c(t, m_i)}), \quad (5)$$

где: $c(t, m_i)$ - частота совместного использования дескриптора t в смысловом значении m_i.

В результате алгоритм поиска по предложенному методу взвешенных дескрипторов можно сформулировать следующим образом:

Шаг 1. Берём в качестве идеала для определения базисов Грёбнера (в нашем случае, концептуальных дескрипторов) текст технического задания на инновационную разработку ИТТ в конкретной предметной области.

Шаг 2. Удаляем из технического задания с помощью стоп-словаря оговоренные выше «помехи».

Шаг 3. Конструируем с помощью словаря-тезауруса данной предметной области концептуальные дескрипторы.

Шаг 4. Упорядочиваем концептуальные дескрипторы в порядке убывания их частоты.

Шаг 5. Определяем диапазон частот наиболее значимых концептуальных дескрипторов (обычно 10-20 дескрипторов).

Шаг 6. Осуществляем запрос и получаем прямоугольную матрицу «дескрипторы – документы» A.

Шаг 7. По формуле (4) упорядочиваем документы в порядке убывания релевантности.

Шаг 8. Проводим на основе латентно-семантического анализа k-аппроксимацию.

Шаг 9. Заносим отобранные документы в базу знаний данной предметной области.

ВЫВОДЫ

В основе предложенного метода взвешенных дескрипторов с использование латентно-семантического анализа лежат принципы факторного анализа, в частности, выявление латентных связей изучаемых явлений или объектов. Кластеризация документов методом взвешенных дескрипторов осуществляется на основе контекстно-зависимых значений лексических единиц при помощи статистической обработки большого объёма текстов. Описанный алгоритм поиска по предложенному методу взвешенных дескрипторов позволяет повысить качество найденной в Интернете информации за счет получения определённых фрагментов информации с учетом семантического анализа текста на естественном языке.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ:

1. Sergeev V.A. Osnovy innovacionnogo proektirovanija – Ul'janovsk : UlGTU 2010. – 246s.
2. Tanenbaum Je., Ujezeroll D. Komp'juternye seti. 5-e izd. — SPb.: Piter, 2012. — 960 s.: il
3. Kuz'min A.V. Zolotareva N.N. Poisk v Internete - Sankt — Peterburg.: Izdatel'stvo NiT, 2011g. 276s.
4. Manning, K. Vvedenie v informacionnyj poisk / K. Manning. – M.: «Vil'jams», 2011.- 200 s.

5. Internet-zavisimoe povedenie (Internet- addictive behavior) : (obzor) : (eview) / V.L.Malygin [i dr.] // Zhurnal nevrologii i psihiatrii imeni S. S. Korsakova. - 2011. - T. 111, № 8. - S. 86-92
6. Bogdanov-Kat'kov, N.V.; Orlov, A.A. Internet: Novejshij spravochnik; M.: Jeksmo, 2012. - 928 c.
7. Denis Shestakov (2011). «Sampling the National Deep Web». Proceedings of the 22nd International Conference on Database and Expert Systems Applications (DEXA), str.331-340.
8. Informatika i IKT. Cvetkova M.S., Velikovich L.S. 3-e izd., ster. - M.: 2012. — 352
9. Dakett, Dzhon Osnovy veb-programmirovanija s ispol'zovaniem HTML, XHTML i CSS / Dzhon Dakett. - M.: Jeksmo, 2015. - 768 c
10. Nekrest'janov I., Pavlova E. Obnaruzhenie strukturnogo podobija HTML-dokumentov. – SPb. Sankt-Peterburgskij gosudarstvennyj universitet, Trudy chetvertoj vserossijskoj konferencii RCDL, 2002. -ss.38-54.
11. Gerdt V.P. Groebner bases and invoiutive methods for algebraic and differential equations // Mathematics and Computers in Modelling, 25, No. 8/9, 1997,pp. 75-90.
12. K. E. Kechedzhy, O. V. Ustenko, V. A. Yampol'ski Rank distributions of words in additive many-step Markov chains and the Zipf Law. – Physical Review E. - 2005. – V.72. - pp. 1-6
13. Wentain Li. Random Texts Exibition Zipf's Law – Like Word Frequency Distribution. Santa Fe institute. NM 87501. - 1992. - V. 38-№6. - pp. 1842-1845
14. Golub Dzh. Matrichnye ischislenija. - M.: Mir. - 1999. -548 c.

Рецензент: д.т.н., проф., Кулаков Ю.А.,
КПИ им. Игоря Сикорского