В статті пропонується латентно-семантичний метод здобуття інформації з інтернет ресурсів, який дозволяє обробляти інформацію на природній мові, а також алгоритм пошуку, що базується на ньому. Основною від'ємністю від існуючих методів є аналіз слів, які зустрічаються у тексті не тільки за частотою, але й враховуючи семантику за рахунок підбору відповідних дескрипторів, що підвищує якість знайденої інформації

Ключові слова: інтернет ресурси, інформаційний пошук, інтелектуальні агенти, дескриптори, закони Зіпфа

В статье предлагается латентно-семантичний метод извлечения информации из интернет ресурсов, позволяющий обрабатывать информацию на естественном языке, а также основанный на нем алгоритм поиска. Основным отличием от существующих методов является анализ встречающихся в тексте слов не только по частоте, но и учитывая семантику, за счет подбора соответствующих дескрипторов, что повышает качество найденной информации

Ключевые слова: интернет ресурсы, информационный поиск, интеллектуальные агенты, дескрипторы, законы Зипфа

УДК 519.21.681

ЛАТЕНТНО-СЕМАНТИЧЕСКИЙ МЕТОД ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ИНТЕРНЕТ РЕСУРСОВ

А. А. Стенин
Доктор технических наук, профессор*
E-mail: alexander.stenin@yandex.ru
Ю. А. Тимошин
Кандидат технических наук, доцент*
E-mail: y.timoshin@kpi.ua
Е. Ю. Мелкумян
Кандидат технических наук*
E-mail: e.melkumyan@ukr.net
В. В. Курбанов
Аспирант*
E-mail: azatotaza@gmail.com
*Кафедра технической кибернетики
Национальный технический университет Украины
«Киевский политехнический институт»
пр. Победы, 37, г. Киев, Украина, 03056

1. Введение

Интернет, в отличие от традиционных Информационно Поисковых Систем (ИПС), имеет следующие особенности [1, 2].

1. Развитие Интернет как информационного хранилища происходило без учёта потребности поиска документа. В результате в Интернет, в отличие от традиционных ИПС, где система хранения документов ориентирована на активный поиск [3], 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, т.к. 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 [4]. 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 [5]. Большая часть информации, хранимой в Интернете, содержится в базах данных (эта часть называется DeepWeb [6]) и недоступно для большинства существующих промышленных ИПС. По оценкам [7], количество документов, хранящихся в базах данных Интернет превышают количество документов хранящихся в промышленных ИПС приблизительно в 500 раз.

2. Анализ литературных данных и постановка задачи исследования

В настоящее время, в смысле автоматизации ИП, 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]. Достижение полной автоматизации в этом вопросе маловероятно, и речь лишь может идти о создании автоматизированных методов и системах извлечения информации из Интернет. Актуальное исследование в области работы со слабо структурированной информацией на основе «интеллектуальных агентов» привели к появлению большого количества альтернативных инструментов их создания. Основные подходы

к решению проблемы извлечения данных из Интернет заимствованы из таких областей как: обработка данных на естественном языке, машинное самообучение, онтология и др. В частности, основной задачей извлечения данных из Web является получение определённых фрагментов информации (поля) из указанных HTML-документов [8, 9]. Эта задача близка к задаче автоматической кластеризации и состоит в поиске разложения HTML-документов $D\{d_1 \ldots d_n\}$ на классы от $C_1 \ldots C_k$, которые содержат документы со схожей структурой. Задания отображения прикладных объектов в точки многомерного пространства состоит в определении базиса признаков $\{e_i\}$, формирующих многомерное пространство, и метода разложения документа по этому базису, то есть вычисление координат $\{w_i\}$.

Для определения координат документа $\{w_i\}$ в пространстве базисных признаков $\{e_i\}$ используются различные подходы. В частности авторами работы [9] представляется использовать подход, популярный при вычислении весов термов в ИПС, использующих векторную модель представления документов.

При этом:

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

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 вводиться энтропийная мера. Однако такой подход, определяющий значимость терма, лишь по частоте – не гарантирует значимость документа по смыслу.

Таким образом, учитывая вышеизложенное, можно определить как основную задачу развития ИПС в Интернете – разработка методов и средств семантического анализа текста на естественном языке.

3. Латентно-семантический метод взвешенных дескрипторов

Как уже упоминалось выше, семантический подход является в настоящее время одним из основных путей совершенствования ИПС, 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 латентно-семантический метод взвешенных дескрипторов, позволяющий извлекать наиболее значимые по смыслу и значению документы, весьма близкие к предметной области конкретной социотехнической системы (СТС).

Метод построен на основе идеи базисов Грёбнера [10], в качестве которых используются статистически построенные концептуальные дескрипторы. Данный метод предполагает, что концептуальные дескрипторы в предложениях имеют низлежащий, «латентный» смысл, который затеняется использованием разных слов. Идеалом при определении базисов Грёбнера будем считать техническое задание на информационное развитие конкретной СТС. Для получения значимых концептуальных дескрипторов воспользуемся законами Джорджа Зипфа, известного американского математика и лингвиста.

Согласно им [11], все создаваемые человеком тексты построены по единым правилам. Зипф предложил, 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).

Значение константы в разных языках различно, но внутри одной языковой группы остаётся неизменным, какой бы текст не был. Так, для английских текстов константа Зипфа $C \approx 1$. Для русских текстов $C \approx 0.06 - 0.07$.

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

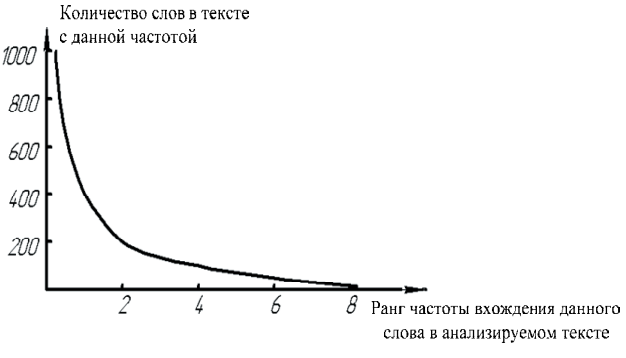

Рис. 1. Кривая Зипфа

В абсолютном масштабе – это гипербола, в логарифмическом масштабе – прямая линия, за исключением нескольких начальных точек.

Объяснение законов Зипфа основано на корреляционных свойствах аддитивных Марковских цепей со ступенчатой функцией памяти [12].

Законы Зипфа универсальны. Им, в частности, отвечают характеристики популярности узлов сети Интернет.

Анализ второго закона Зипфа показывает, что наиболее значимые слова, а, следовательно, и дескрипторы лежат в средней части кривой Зипфа. 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.

Во многом определение диапазона зависит от корректного составления двух специальных словарей – тезауруса и стоп-словаря. Тезаурус данной предметной области даёт возможность корректно определить набор концептуальных дескрипторов и технического задания и наиболее значимых смыс-

ловых понятий данной предметной области. Стоп-словарь отсекает «помехи» в виде «лишних» слов, т.е. для русского языка – это частицы, предлоги, местоимения и др.

Выбор количества дескрипторов определяется заданным разработчиком частотным диапазоном выбранных из словаря тезауруса и скорректированных техническим заданием дескрипторов. Частотный диапазон определяется анализом частоты их появления в техническом задании.

Пусть первоначально 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, $i=\overline{1,n}, j=\overline{1,N}$.

Т.к. 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 коллекцией (набором) документов и терминами, в них встречающимися, составляющий некоторые факторы всем документам и терминам.

В основе метода ЛСА лежат принципы факторного анализа, в частности, выявление латентных связей изучаемых явлений или объектов [13]. При кластеризации документов ЛСА использует для извлечения контекстно-зависимых значений лексических единиц при помощи статистической обработки большого объёма текстов.

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.

Основная идея k-аппроксимации латентно-семантического подхода к обработке матрицы A состоит в замене её 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 [13], прямоугольная вещественная 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, т.е. 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.

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

Как правило, выборка зависит от поставленной задачи и выбирается эмпирически.

Для выбора k сконструируем критерий релевантности следующего вида:

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

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 значимости можно определить на основе экспертных оценок смысловой значимости i-го дескриптора в техническом задании, как идеала будущего набора документов. Пусть число экспертов будет m (табл. 1).

Таблица 1

Ранжирование дескрипторов

Дескрипторы / Эксперты	1	2	...	N
1	λ_{11}	λ_{12}	...	λ_{1n}
2	λ_{21}	λ_{22}	...	λ_{2n}
...	...	...	...	...
m	λ_{m1}	λ_{m2}	...	λ_{mn}

Каждый эксперт производит ранжирование по смысловому значению дескрипторов на основе упорядочивания. При ранжировании эксперт должен расположить дескрипторы в порядке, который ему представляется наиболее оптимальным с точки зрения их смысловой значимости, и приписать каждому из них числа натурального ряда – ранги (показатели, характеризующие порядковое место оцениваемых дескрипторов). При этом наиболее значимому дескриптору приписывается первый ранг, а наименее значимому – последний, т.е. в нашем случае n. Если эксперт не в состоянии указать порядок следования двух или нескольких дескрипторов, либо он присваивает разным дескрипторам один и тот же ранг, то по требованию, что порядковая шкала должна удовлетворять условию, равенства числа рангов n числу ранжируемых дескрипторов, дескрипторам присваиваются стандартные ранги, представляющие собой среднее суммы мест, поделённых между собой дескрипторами с одинаковыми рангами.

В случае согласованного мнения экспертов можно определить коэффициенты относительной значимости $\tilde{\alpha}_{ij}$ как:

$$\tilde{\alpha}_{ij} = \frac{\sum_{j=1}^{m} \lambda_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{n} \lambda_{ij}}. \qquad (5)$$

Из (5) следует, что смысловой коэффициент значимости i-го дескриптора будет равен:

$$\alpha_i = \frac{\sum_{j=1}^{m} \tilde{\alpha}_{ij}}{m}. \qquad (6)$$

Отсюда формулы (4) и (6) позволяют правильно сформировать матрицу A, а затем выбрать k наибольших сингулярных чисел.

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

Шаг 1. Берём в качестве идеала текст технического задания на инновационное развитие СТС в конкретной предметной области.

Шаг 2. Удаляем с помощью стоп-словаря «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 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 ЛСА k-аппроксимацию.

Шаг 9. Заносим отобранные документы в БД.

4. Выводы

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

Литература

1. Козлов, Д. Д. ИПС в Интернет: текущее состояние и пути развития [Текст] / Д. Д. Козлов. – М.:МГУ. – 2000. – 28 с.
2. Ландэ, Д. В. Поиск знаний в Internet [Текст] / Д. В. Ландэ. – М.: Диалектика. – 2005. – 28 с.
3. Мидоу, Ч. Ч. Анализ информационно-поисковых систем [Текст] / Ч. Ч. Мидоу. – М.:Мир. – 1970.
4. Lawrence, S. Accessibility of Information on the Web [Текст] / S. Lawrence, C. Giles // Nature. – 1999. – vol. 400 – C. 107-109.
5. Hermans, B. Intelligent Software Agents on the Internet [Електронний ресурс] / B. Hermans. – 1996. – 89 с. – Режим доступу: \www/ URL: http://www.hermans.org/agents.
6. Bergman, K. The Deep Web: Surfacing Hidden Value, BrightPlanet.com LLC [Електронний ресурс] / K. Bergman. – Режим доступу: \www/ URL: http://www.completeplanet.com/Tutorials/DeepWeb/index.asp.
7. Inktomi Corp., Web Surpasses One Billion Documents, press release issued January 18, 2000 [Електронний ресурс]. – Режим доступу: \www/ URL: http://www.inktomi.com/new/press/billion.html.
8. Методы и средства извлечения слабоструктурированных схем из документов в HTML и конвертирования HTML документов в их XMLпредставление [Електронний ресурс]. – Режим доступу: \www/ URL: http://synthesis.ipi.ac.ru/syntesis/projects/XMLBIS/html2xml.html.
9. Некрестьянов, И. Обнаружение структурного подобия HTML-документов [Електронний ресурс]/ И. Некрестьянов, Е. Павлова. – СпбГУ, 2002. – С. 38-54. – Режим доступу: \www/ URL: http://meta.math.spbu.ru.
10. Gerdt, V. P. Computer Algebra and Constrained Dynamics [Текст] / V. P. Gerdt // Problem of Modern Physics. – 2000. – JINR D2-99-263. – C. 164-171.
11. Kechedzhy, K. E. Rank distributions of words in additive many-step Markov chains and the Zipf Law [Текст] / K. E. Kechedzhy, O. V. Ustenko, V. A. Yampol'ski // Arxiv LANL. – 2004. – Phys.Rev.E. – 2005. – V 72. – pp. 1-6.
12. Wentain, Li. Random Texts Exibition Zipf's Law – Like Word Frequency Distribution. [Текст] / Li. Wentain // Santa Fe institute. NM 87501. – 1992. – V. 38-№6. – C. 1842-1845.
13. Голуб, Дж. Матричные исчисления [Текст] / Дж. Голуб, И. Ван Лоун. М.: Мир. – 1999.