P. V. Val

SHORT-TERM FORECASTING OF MINING ENTERPRISE ENERGY DEMAND BY UNIVARIATE METHODS

In this paper the author performs analysis of short-term forecasting quality of mining enterprise energy demand, using the popular univariate forecasting methods (seasonal ARIMA, Holt-Winters method, and naive method).

Keywords: energy consumption, forecasting, ARIMA, Holt–Winters method, naive method.

УДК 62-506.1

П. В. Зеленков, Г. А. Сидорова

МОДИФИЦИРОВАННЫЙ АЛГОРИТМ HITS*

Показана проблема современных поисковых систем, связанная с ранжированием документов. Для решения данной проблемы в процессе поиска и обработки информации предлагается использовать модифицированный алгоритм HITS. Данный подход помогает решать проблемы поиска, определения релевантности найденной информации, а также производить ранжирование отклика системы.

Ключевые слова: HITS, ранжирование, обработка информации, поиск информации.

В настоящее время при создании и развитии технологий сбора и обработки информации основное внимание удаляется развитию существующих технологий, нацеленных на анализ баз данных поисковых сервисов сети Интернет, и развитию алгоритмов ранжирования [1; 2]. Однако если встает вопрос об организации подобных процедур в рамках локальных корпоративных систем, то возникает проблема в анализе информации и ее взаимосвязей на локальном уровне.

На сегодняшний день существует множество алгоритмов ранжирования информации в поисковых системах сети Интернет [3]. Один из самых распространенных – это алгоритм Клейнберга, для которого создано несколько модификаций. Наиболее значимым является метод HITS, который заключается в присвоении каждому документу в веб-множестве некоторых значений, которые называются весами документа. Существует два вида таких весов: *a* (authority) – вес авторитетного документа и *h* (hub) – вес хаб-документа. Авторитетный документ – это документ, соответствующий запросу пользователя, имеющий больший удельный вес среди документов данной тематики, т. е. большее число документов ссылается на данный документ. Хаб-документ – это документ, содержащий много ссылок на авторитетные документы. Соответственно, для каждой страницы рассчитывается не один, а два веса. Такой подход обусловлен наличием в Сети большого числа сообществ, т. е. наборов страниц близкой тематики, которые весьма сильно связаны друг с другом ссылками. Исходя из значений весов, происходит формирование множества поиска и его ранжирование по релевантности.

Такой подход очень удобен, так как позволяет находить больше документов, соответствующих заданной тематике. Однако у него есть и недостатки, которые естественным образом вытекают из достоинств: во множество найденных документов может попасть большое количество страниц с низким коэффициентом релевантности, которые, тем не менее, имеют много ссылок друг на друга, и именно им будут присвоены наивысшие ранги. Это явление называется смещением тематики (diffusion, drift). Обычно оно происходит в направлении более широкой предметной области (или лучше представленной в Сети). Для решения этой проблемы Клейнберг предложил использовать анализ содержимого страниц, но оценивать не отдельные страницы, а разные сообщества целиком.

Описание модифицированного метода. Модифицированный метод может быть полезен для поиска как в корпоративных информационно-управляющих системах, так и в локальных и глобальных сетях. Основа метода – избирательный поиск не по всему веб-пространству, а по документам, принадлежащим внутренней сети.

Его очень удобно использовать в организациях, специализирующихся на узкой тематике и имеющих обширную базу данных. Разработанная на базе этого метода поисковая система будет не только обрабатывать нужные документы, но и производить пополнение внутренней базы документами смежной тематики, найденными в Сети.

*Работа выполнена в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг.

Рассмотрим предлагаемый метод подробнее. Для этого введем обозначения: $\{W\}$ – множество документов Сети; $\{K\}$ – множество документов сети (корпоративной, локальной).

Каждое множество содержит конечное число элементов. Пусть W содержит N элементов, а K содержит M элементов, причем $M < N$. Тогда каждый документ можно представить в одном из следующих видов:

$d_i \in W, 1 \le i \le N$, если документ принадлежит Сети;

$d_j \in K, 1 \le j \le M$, если документ принадлежит локальной сети.

Особо ценными для корпоративных сетей являются документы вида $d_j \in K / W$, т. е. внутренние документы корпоративной сети.

Поиск документов производится по объединению этих множеств, т. е. по $W \bigcup K$ (рис. 1).

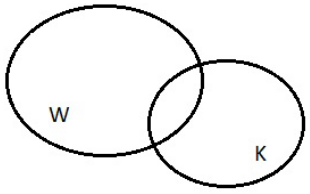

Рис. 1. Множество поиска

На множестве поиска строим граф G, называемый графом поиска. Для этого выделим множество документов R, которое представляет собой множество вершин графа G, и множество A ссылок между документами, которое является множеством ребер графа G. Таким образом, граф $G[R, A]$ – граф поиска, в котором можно выделить подграф $G'[R', A']$, где R' – множество документов, принадлежащих локальной сети, т. е. элементами которого являются документы вида $d_j \in K, 1 \le j \le M$; A' – множество входящих и исходящих ссылок, принадлежащих документам из R' (рис. 2).

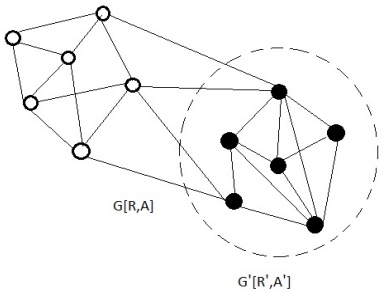

Рис. 2. Граф поиска

Ссылки в документах встречаются двух типов – входящие и исходящие. Количество входящих ссылок определяет авторитетный документ, количество исходящих ссылок определяет хаб-документ. Рассчитаем формулу для вычисления весов. Если вершина U является хаб-документом, а вершина V – авторитетный документ, то $U \to V;\ U, V \in G[R, A]$. Пусть на документ V ссылается n хаб-документов, тогда $U_i \to V;\ U_i, V \in G[R, A],\ i = 1...n$. Аналогично формуле из метода HITS, вес авторитетного документа V вычисляется следующим образом:

$$A(V) = a_V \sum_{\substack{i=1 \\ U_i \to V}}^{n} H(U_i),$$

где $H(U_i)$ – веса хаб-документов U_i, $i = 1 \ldots n$; a_V – индекс сети, который мы приняли равным 1, в случае если документ V не принадлежит внутренней сети, 2 – если документ V принадлежит внутренней сети, т. е.

$$a_V = \begin{cases} 1, & \text{если } V \in W / K, \\ 2, & \text{если } V \in K. \end{cases}$$

Таким образом, веса авторитетных документов, принадлежащих внутренней сети, будут в два раза выше, чем веса внешних авторитетных документов. Иными словами, релевантность документов сети будет выше, что позволит экономить внешний трафик и значительно увеличит скорость работы корпоративной сети.

Веса хаб-документов вычисляются через веса авторитетных документов. Пусть хаб-документ U ссылается на m авторитетных документов, т. е. $U \to V_j;\ U, V_j \in G[R, A],\ j = 1...m$.

Тогда вес документа U вычисляется по формуле

$$H(U) = a_V \sum_{\substack{j=1 \\ U \to V_j}}^{m} A(V_j),$$

где $A(V_j)$ – веса авторитетных документов $V_j, j = 1...m$; a_V – индекс сети.

Для каждого документа рассчитываются два веса: вес документа как первоисточника (авторитетный вес) и вес документа как посредника (хаб-вес). Документы, принадлежащие внутренней сети, после ранжирования будут иметь приоритетный ранг по сравнению с документами Сети, что позволит пользователю просматривать их в первую очередь.

Этапы работы модифицированного алгоритма. Модифицированный алгоритм HITS работает аналогично алгоритму, созданному Клейнбергом.

На первом этапе происходит составление корневого множества документов, релевантных запросу – Root Set. Для этого производится поиск по ключевым словам в базе данных информационно-управляющей системы и из ответа извлекается k первых результатов. Возьмем $k \le 200$ и рассмотрим процесс формирования Root Set (рис. 3).

На втором этапе к страницам из Root Set добавляются их ближайшие соседи, т. е. те страницы, на которые ссылаются страницы из Root Set, и те, которые сами имеют ссылки на страницы Root Set. Для поиска последних также используется поисковая система, причем берется не более d входящих ссылок на одну страницу. Так строится Base Set.

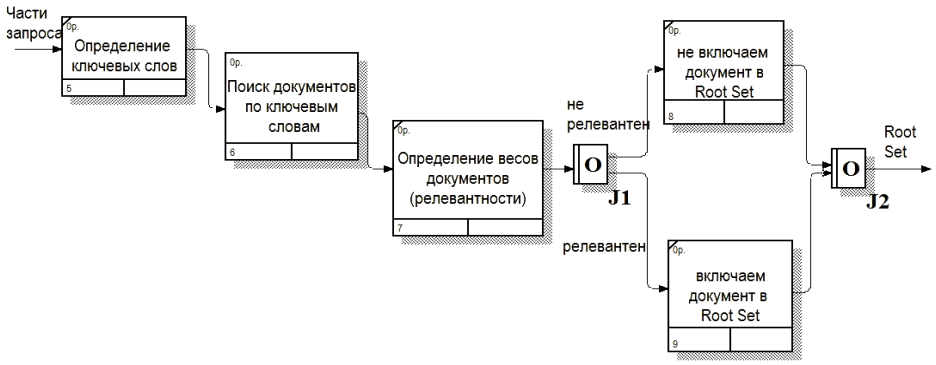

Рис. 3. Формирование Root Set

На третьем этапе к множеству Base Set применяется метод, приведенный выше, т. е. происходит построение графа поиска и вычисление весов документов. В качестве модификации предлагаемого метода выступает индекс сети a_V. Зачастую в Сети встречаются страницы, имеющие высокие авторитетные и хаб-веса, но содержащие мало информации по искомой тематике. Это могут быть рекламные страницы либо страницы с искусственно завышенным рейтингом. Модифицированный метод помогает избежать попадания таких страниц в список найденных документов.

Ранжирование множества документов Base Set с помощью предложенного метода показано на рис. 4. Вычисление весов документов (процессы 16 и 17), согласно их принадлежности сети, происходит по следующим формулам:

$$a_V = 2, \quad A(V) = 2 \sum_{\substack{i=1 \\ U_i \to V}}^{n} H(U_i), \quad H(U) = 2 \sum_{\substack{j=1 \\ U \to V_j}}^{m} A(V_j);$$

$$a_V = 1, \quad A(V) = \sum_{\substack{i=1 \\ U_i \to V}}^{n} H(U_i), \quad H(U) = \sum_{\substack{j=1 \\ U \to V_j}}^{m} A(V_j).$$

Таким образом, веса документов корпоративной сети будут иметь приоритетный коэффициент релевантности относительно документов из Сети. Это позволит пользователю находить более информативные документы, которые из-за невысокого рейтинга игнорируются другими поисковыми системами.

Модифицированный метод предназначен для обработки информации в корпоративных информационно-управляющих системах. На данный момент он используется в разработке демонстрационного прототипа корпоративной информационно-управляющей системы. Первые экспериментальные результаты работы модифицированного метода ожидаются в сентябре 2011 г. По теоретическим данным, метод снижает загруженность корпоративного интернет-трафика при поиске и обработке информации. Снижение внешнего интернет-трафика достигается благодаря формированию статистики наиболее часто используемых документов и пополнению корпоративной базы данных согласно запросам пользователя. Кроме того, данный метод позволяет повысить ранг внутренних корпоративных документов по сравнению с внешними, а тем самым – достоверность и, как следствие, качество результирующей информационной выборки.

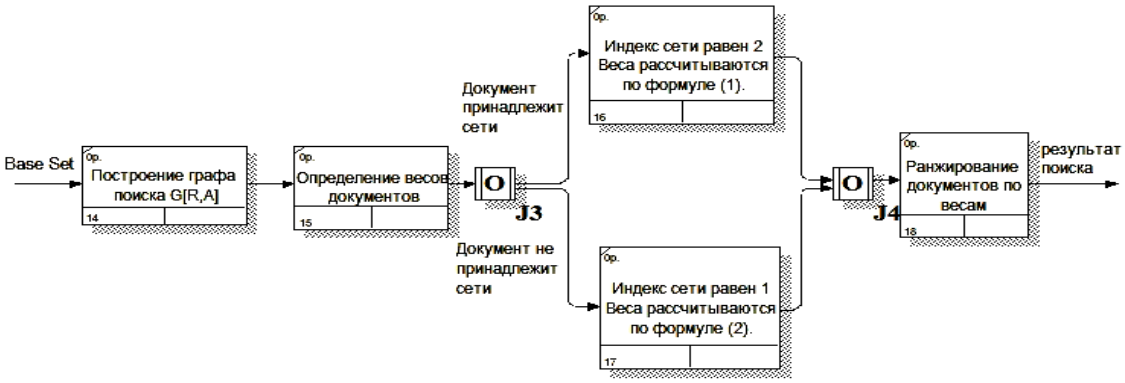

Рис. 4. Ранжирование

Библиографические ссылки

1. Information search module based on multilinguistic thesauruses / P. V. Zelenkov, M. V. Karaseva, E. P. Bachurina, V. V. Brezitskaya // Вестник СибГАУ. Вып. 1(27). 2010. С. 89–91.

2. Модуль обработки информационных запросов пользователей в сеть Интернет для корпоративных информационно-управляющих систем / П. В. Зеленков, М. А. Селиванова, В. В. Брезицкая, А. П. Хохлов // Вестник СибГАУ. Вып. 3(24). 2009. С. 69–74.

3. System for processing highly specialized information in distributed networks / P. V. Zelenkov, V. V. Brezitskaya, E. P. Bachurina et al. // Вестник СибГАУ. Вып. 5(26). 2009. С. 40–42.

P. V. Zelenkov, G. A. Sidorova

MODIFICATED HITS ALGORITHM

In this paper the problem of modern search systems connected with documents ranking is shown. To solve this problem it is proposed to use the modified algorithm HITS in the process of searching and processing information. This approach helps to solve the problems of search, relevance determination of the information and also to rank system response.

Keywords: HITS, ranking, information processing, information search.

УДК 519.688

А. В. Ищенко, И. В. Киреев

ФРАКТАЛЬНЫЙ АЛГОРИТМ ПОСТРОЕНИЯ ДВУМЕРНЫХ ВЛОЖЕННЫХ СЕТОК*

Предложен алгоритм построения последовательности вложенных сеток для двумерных многосвязных областей, сочетающий в себе достоинства алгоритма шаблонов и метода бисекции.

Ключевые слова: триангуляция, фрактал, многосеточные методы.

Построению двумерных сеток посвящено множество публикаций, обзор которых можно найти в [1–3]. Тем не менее, появление любого нового алгоритма вызывает большой интерес, поскольку триангуляция является одним из основных этапов численного моделирования в механике сплошных сред. В данной статье предложен алгоритм разбиения двумерных многосвязных областей на треугольники, идея которого подсказана алгоритмом построения салфетки Серпинского, представляющей собой один из классических примеров фрактальной геометрии [4].

В основе рассматриваемого алгоритма триангуляции, результатом работы которого будет последовательность неравномерных вложенных друг в друга сеток, лежит единственность представления прямоугольного равнобедренного треугольника в виде объединения двух прямоугольных равнобедренных треугольников, полученных из исходного бисекцией (рис. 1).

Как и в методе граничной коррекции [2], исходную многосвязную двумерную область Ω помещаем в прямоугольник R, на котором определена характеристическая функция $\chi(M)$ области $\Omega \subseteq R$:

$$\chi(M) = 1 \Leftrightarrow M \in \Omega \text{ и } \chi(M) = 0 \Leftrightarrow M \notin \Omega. \qquad (1)$$

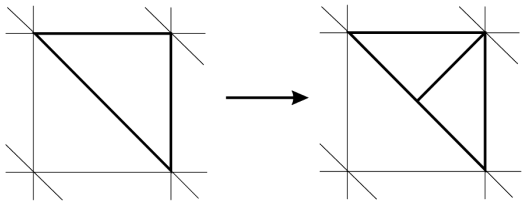

Рис. 1. Шаблон для прямоугольного треугольника

Аналогично методу шаблонов [3], в котором в качестве базового элемента взят равнобедренный прямоугольный треугольник и шаблоном для которого является указанное выше представление в виде объединения двух равных треугольников, триангулируем объемлющий прямоугольник R, после чего при помощи граничной коррекции [3] получим сетку для области Ω.

Для описания алгоритма триангуляции объемлющего прямоугольника R, который изначально разбит на небольшое количество равнобедренных прямоугольных треугольников, удобно ввести понятие уровня вложенности треугольника [5]. Треугольникам начального разбиения прямоугольника R приписываем нулевой уровень вложенности.

*Работа выполнена при финансовой поддержке РФФИ «Вычислительные технологии для расчета течений несжимаемой жидкости» (проект № 08-01-00621-а).