the construction of main gas pipelines. Ballasting, ensuring stability of gas pipelines at project marks]. Moscow, Information and Advertising Center Gazprom, 1996, pp. 106–149.

6. *STO Gazprom 2-2.2-578-2011. Sredstva ballastirovki i zakrepleniya gazoprovodov v proektnom polozhenii. Tipovye metodiki ispytaniy* [Means for ballasting and fastening gas pipelines in project position. Standard testing methods]. Moscow, Gazprom expo, 2011. 60 p.

7. Khallyev N. Kh., Reshetnikov A. D., Budzulyak B. V. et al. *Kapitalnyy remont lineynoy chasti magistralnykh gazonefteprovodov* [Overhaul of the linear part of main gas and oil pipelines]. Moscow, Max Press, 2011. 448 p.

УДК 004.91

МАТЕМАТИЧЕСКИЕ МОДЕЛИ ИНФОРМАЦИОННОГО ПОИСКА WEB-РЕСУРСОВ

Кузнецов Михаил Андреевич, кандидат технических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, mara122@mail.ru

Нгуен Тан Там, магистрант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, tantamvn@gmail.com

Реализация поисковой системы для нахождения web-ресурсов требует построения адекватной математической модели. Подавляющее большинство разработанных моделей ориентировано на текстовый поиск. Модель должна обеспечивать высокую скорость обработки поисковых запросов, вести качественную оценку релевантности и ранжируемости результатов. Существующие системы, такие как GOOGLE, YAHOO, BING и т.д., используют подобные математические модели. Каждая реализация имеет отличия, характеризующиеся преимуществами и недостатками. Несмотря на принципиальные особенности в реализации моделей, можно выделить несколько базовых подходов. Каждый подход использует определенные принципы обработки и представления текста для поиска. Статья посвящена рассмотрению особенностей базовых математических моделей, лежащих в основе построения существующих поисковых систем. Рассматриваются модели на основе множеств: векторные, вероятностные и ссылочные.

Ключевые слова: поисковая система, релевантность, ранжирование результатов поиска, ссылочное ранжирование, web-поиск, ключевые слова, термы, лексема, лексический анализ, поисковый запрос

MATHEMATICAL MODELS OF WEB RESOURCES SEARCH

Kuznetsov Mikhail A., Ph.D. (Engineering), Volgograd State Technical University, 28, Lenin av., Volgograd, 400005, Russian Federation, e-mail: mara122@mail.ru

Nguen Tan Tam, undergraduate student, Volgograd State Technical University, 28, Lenin av., Volgograd, 400005, Russian Federation, e-mail: tantamvn@gmail.com

The implementation of a search engine to find web resources requires the construction of the adequate mathematical model. The vast majority of the developed models are focused on text search. Model should provide the high-speed processing of search queries, conduct a qualitative assessment of the relevance and a ranking-orientability results. Existing systems such as GOOGLE, YAHOO, BING, etc. are using such mathematical models. Each implementation is different and characterized by benefits and drawbacks. Despite the fundamentally especially in the implementation of models, there are several basic approaches. Each approach uses certain principles of processing and presentation of text to the search. The article considers the basic features of the mathematical models that underlie the construction of existing search engines. The article contain model based on sets, vectors, probability and references.

Keywords: search engine, relevance, result ranking, PageRank, reference ranking, web search, keywords, term, lexeme, lexical analysis, search request

Огромный объем информационных ресурсов web выводит задачу поиска необходимых данных на качественно новый уровень. Задачу усложняют факторы изменчивости ресурсов и постоянного роста их количества. Для эффективной работы поисковых систем необходимо выполнение нескольких требований. Важнейшими из них являются скорость обработки поискового запроса, релевантность результатов поиска и возможность грамотного ранжирования найденных документов.

Существующие мощные и наиболее известные крупные информационно-поисковые системы, такие как Google, Yahoo, Bing, Яndex, Rambler и др., охватывают миллиарды вебдокументов. Такие системы отличаются друг от друга специальными алгоритмами, позволяющими обеспечить качественный и быстрый поиск. Но все эти алгоритмы являются модификациями основных подходов – моделей поиска.

Модель поиска – это некоторое упрощение реальности, на основании которого получается математическая формула и правила применения этой формулы к документам. Формула и правила позволяют системе принять решение, какой документ считать соответствующим поисковому запросу и как ранжировать множество найденных документов. В основу традиционных методов положены три главных подхода.

Первый подход базируется на теории множеств. В качестве разновидностей данного подхода можно выделить следующие виды: булевская, расширенная булевская модель и нечеткие множества.

Второй подход основывается на векторной алгебре. Этот подход можно представить в виде векторной, обобщенной векторной, латентно-семантической и нейросетевой моделях.

Третий подход происходит из теории вероятностей – вероятностная модель [4, с. 5].

Классические модели информационного поиска рассматривают документы как множества представляющих эти документы ключевых слов, в дальнейшем называемых термами. Терм (англ. – *term*) является просто словом, семантика которого помогает описать основное содержание документа.

Любая модель информационного поиска представляется в виде следующих составляющих:

1. Формат представления документа.

2. Формат представления запроса – формализованный способ выражения информационных потребностей пользователя системы.

3. Функция соответствия документа запросу – степень соответствия запроса и найденного документа (релевантность).

Пусть i – индекс терма t_i из словаря T ($i = 1, ..., M$), $d^{(j)}$ – документ j, принадлежащий множеству документов D, а $w^{(i,j)} \geq 0$ – вес, ассоциированный с парой $(t_i, d^{(j)})$. Для каждого терма t_i, который не входит в документ $d^{(j)}$, его вес равен нулю: $w^{(i,j)} = 0$.

Булевская модель основывается на теории множеств и математической логике. Документы и запросы представляются в виде множества термов – ключевых слов. Каждый терм представлен как булева переменная: 0 (терм из запроса не присутствует в документе) или 1 (терм из запроса присутствует в документе). При этом весовые значения терма в документе принимают лишь два значения: $w^{(i,j)} \in \{0,1\}$.

В булевских моделях поиска пользователь может формулировать запрос в виде булевского выражения, используя для этого операторы И, ИЛИ, НЕТ. Известно, что любое логическое выражение можно представить дизъюнкцией некоторых выражений, соединенных между собой операцией конъюнкции (дизъюнктивной нормальной формой – ДНФ). Можно записать:

$$q \equiv d_{dnf} = \bigvee_{i=1...N} q_{cc}^{(i)},$$

где q – запрос; $q_{cc}^{(i)}$ – i -я конъюнктивная компонента формы запроса q_{dnf}. Тогда мера близости документа $d^{(j)}$ и запроса q – $sim(d^{(j)}, q)$ (от англ. *similarity* – близость) в булевой модели определяется выражением 1 [3, с. 31]:

$$sim(d^{(j)}.q) = \begin{cases} 1, если\ \exists q_{cc}^{(i)} : \left(q_{cc}^{(i)} \in q_{dnf}\right) \wedge \left(\forall k, g_k\left(q_{cc}^{(i)}\right) = g_k(d^{(j)})\right) \\ 0, иначе \end{cases}, \quad (1)$$

где g_k – инверсная функция, соответствующая индексу терма t_k, которая определяется следующим образом: $g_k(d^{(j)})=w_k^j$, т.е. $sim(d^{(j)}, q) = 1$, если существует такая конъюнктивная компонента $q_{cc}^{(i)}$, входящая в дизъюнктивную нормальную форму q_{dnf}, что инверсная функция каждого терма к данной конъюнктивной компоненты совпадает с этой же инверсной функцией для документа $d^{(j)}$. В противном случае $sim(d^{(j)}, q)$ оказывается равной 0.

Таким образом, если $sim(d^{(j)}, q) = 1$, то в соответствии с булевой моделью документ $d^{(j)}$ считается релевантным запросу q. В противном случае документ не является релевантным.

Одним из несомненных достоинств булевской модели поиска является простота ее реализации. Главными недостатками считаются [1, с. 105]:

- невысокая эффективность поиска, отсутствие контекстных операторов, отсутствие возможности ранжирования найденных документов по степени релевантности, поскольку отсутствуют критерии ее оценки;
- сложность использования – далеко не каждый пользователь может свободно оперировать булевскими операторами при формулировке своих запросов.

Основной недостаток классической булевой модели связан с отсутствием весовых значений термов в поисковом запросе, а значит, и нивелированием значимости отдельных термов. Это приводит к невозможности ранжирования результатов поиска по уровню их соответствия информационным запросам. Для того чтобы устранить этот недостаток и вместе с тем использовать вычислительные преимущества булевой модели, предложено несколько вариантов расширенных булевских моделей. В этих моделях вводятся специальные обобщения булевских операторов на основе нечетких множеств, позволяющие учитывать меру соответствия документа выражению запроса.

Векторная модель является классическим представителем класса алгебраических моделей. В рамках этой модели документы и запросы описываются в виде векторов в многомерном пространстве термов. Каждому терму, использующемуся в документе, ставится в соответствие весовое значение. Значение определяется на основе статистической информации о количестве появлений терма в рассматриваемом документе и во всем документальном массиве. В векторной модели не предусмотрено использование логических операций в запросах. Для оценки близости запроса и документа используется скалярное произведение соответствующих векторов запроса и документа.

Близость документа $d^{(j)}$ к запросу q рассматривается как скалярное произведение информационных векторов, представленных весовыми значениями термов $\overline{d^j} = (w_1^{(j)}, w_2^{(j)},..., w_n^{(j)})$ и $\overline{q} = (w_1^q, w_2^q,..., w_n^q)$. При этом вес отдельных термов можно вычислять разными способами. Один из возможных простейших подходов основан на использовании в качестве веса терма $w_i^{(j)}$ нормализованной частоты $freq_i^{(j)}$ встречаемости терма в данном документе с учетом частоты нахождения данного терма в других документах коллекции. Этот способ называют учетом дискриминационной силы терма (см. формулу 2):

$$w_i^{(j)} = freq_i^{(j)} \cdot \log\left(\frac{N}{n_i}\right), \quad (2)$$

где n_i – количество документов, в которых используется терм t_i, а N – общее количество документов в массиве. Например, если некоторое слово встречается в каждом документе массива, то его использование в запросе, очевидно, бесполезно. Соответственно, в этом случае $n_t = N$ и, следовательно, $w_i^{(j)} = freq_i^{(j)} \cdot \log\left(\frac{N}{N}\right) = 0.$

Такой метод взвешивания термов имеет стандартное обозначение – TF* IDF, где TF (от англ. *Term Frequency* – частота термина) указывает на частоту появления терма в документе, а IDF (от англ. *Inverse Document Frequency* – обратная частота документа) – на величину, обратную количеству документов в массиве, содержащих данный терм.

Для определения тематической близости документа и запроса, в этой модели используется простое скалярное произведение *sim(dj,q)*, которое соответствует косинусу угла между векторами $\overline{d^j}$ и $\overline{q}$. Мерой близости документа $d^{(j)}$ и запроса q является величина, рассчитываемая по формуле 3 [3, с. 40]:

$$\sin\ (d_j,\ q) = \frac{\overline{d^{(j)}}}{\left|\overline{d^{(j)}}\right|} \cdot \frac{\overline{\cdot q}}{\left|\overline{q}\right|} = \frac{\sum_{i=1}^{n} w^{(j)} w_i^q}{\sqrt{\sum_{i-1}^{n} (w_i^{(j)})^2} \sqrt{\sum_{i-1}^{n} (w_i^q)^2}}. \quad (3)$$

Векторная модель наиболее часто используется на практике, так как она реализуется довольно просто, обеспечивает эффективность поиска и ранжирования. Кроме этого, векторно-пространственная модель обеспечивает поисковым системам возможность простой реализации режима поиска подобных документов. Ведь каждый документ может рассматриваться как запрос. Но вместе с тем векторно-пространственная модель связана с расчетом массивов высокой размерности и в каноническом виде малопригодна для обработки больших массивов данных.

Фундаментом *вероятностной модели* поиска выступает теория вероятностей. Релевантность в этой модели рассматривается как вероятность того, что данный документ может оказаться интересным пользователю. При этом подразумевается наличие уже существующего первоначального набора релевантных документов (учебная выборка), выбранных пользователем или полученных автоматически при каком-нибудь упрощенном предположении. Вероятность оказаться релевантным для каждого следующего документа рассчитывается на основании соотношения встречаемости терминов в релевантном наборе и в остальной, «нерелевантной» части коллекции.

В данной модели поиска вероятность того, что документ релевантен запросу основывается на предположении, что термы запроса по-разному распределены среди релевантных и нерелевантных документов. При этом используются формулы расчета вероятности, базирующиеся на теореме Байеса. В соответствии с теоремой Байеса, по некоторой функции вероятностей получим конечную форму, которая оценивает уровень вероятности релевантности для каждого документа из учебной выборки, называемую поисковым статусом (см. формулу 4 [3, с. 45]):

$$SV = \sum_{t_i \in q \cap d} SV = \sum_{t_i \in q \cap d} \log \frac{rel_i (nrel - nrel_i)}{nrel_i (rel - rel_i)}, \quad (4)$$

где rel_i – количество релевантных документов, которое содержит терм с индексом i; $nrel_i$ – соответственно, количество нерелевантных документов; d – учебная выборка документа рассматривается как множество слов; q – множество слов, входящих в запрос, $q \cap d$ означает множество общих термов в запросе и документе.

По данным экспертной оценки релевантности запроса для документов из учебной выборки рассчитываются значения rel_i и $nrel_i$, а также экспоненты от соответствующей со-

ставляющей поискового статуса релевантности $exp(SV_i)$ для каждого терма из запроса. Для дополнительных документов (документов, оцениваемых не экспертами) значение поискового статуса SV релевантности рассчитывается в соответствии с вышеприведенной формулой. Таким образом, имея образцы релевантных документов, можно получить выборку дополнительных документов и ранжировать их в соответствии с проведенным расчетом.

Вероятностные модели обладают некоторым теоретическим преимуществом, они предлагают наиболее естественный способ формально описать проблему информационного поиска и при имеющейся информации дают наилучшие предсказания релевантности. На практике они так и не получили большого распространения, так как вероятностная модель характеризуется низкой вычислительной масштабируемостью (т.е. резким снижением эффективности при росте объемов данных) и необходимостью постоянного обучения системы.

Рассмотренные модели могут применяться на практике и в каноническом виде. Однако у них есть общий недостаток, обусловленный предположением, что содержание документа определяется множеством слов, которые входят в него без учета взаимосвязей между термами. Смысл текста не анализируется вообще. Теоретически можно построить документ с бессмысленным сочетанием набора термов, который будет иметь высокую степень релевантности какому-либо запросу. Вряд ли это то, что желает получить пользователь поисковой системы. Поэтому существуют модели поиска, анализирующие смысл, например, семантические. В рамках подобных моделей делаются попытки организации смыслового поиска за счет анализа грамматики текста, использования баз знаний, тезаурусов, онтологий. Все эти модели реализуют учет семантической связи между отдельными словами и их группами. Вместе с тем эффективность систем, базирующихся на таких подходах, пока остается невысокой [3, с. 29]. На практике чаще всего используются гибридные подходы, в которых объединены возможности булевой и векторной моделей и зачастую добавлены оригинальные методы семантической обработки информации.

Рассмотрение математических моделей, учитывающих только текстовое содержимое документов, было бы не полным, так как современные поисковые системы анализируют также структуру гипертекстовых ссылок. Эта информация помогает учесть авторитетность опубликованных ресурсов. Показатель авторитетности очень важен. Интернет предоставляет десятки и даже сотни тысяч релевантных запросу документов. Ранжирование позволяет установить границу между теоретически и практически найденными документами (рядовые пользователи просматривают только верхние 10–30 найденных документов). При использовании любых методов анализа текстов документа и запроса (т.е. любого механизма оценки релевантности) выдача результата требует дополнительного учета важности найденных ресурсов. При расчете важности документа может учитываться несколько видов факторов [2, с. 43]. Наиболее используемой на практике стала модель PageRank. Она представляется формулой 5:

$$PR_\alpha = \frac{(1-d)}{N} + d\sum_{i-1}^{n}\frac{PR_i}{C_i}, \qquad (5)$$

где PR_α – PageRank рассматриваемой страницы; d – коэффициент затухания; N – общее количество документов; PR_i – PageRank i-й страницы, ссылающейся на рассматриваемую страницу; C_i – общее число ссылок на i-й странице.

В основу вычисления PageRank положена вероятностная модель блуждающего по документам сети пользователя. Вероятность того, что пользователь посетит конкретный документ, принимается за важность (ранг) документа. В моделях поиска, использующих PageRank, релевантные документы сортируются на основе данного показателя или каким-либо образом учитывают его при сортировке.

Важным достоинством PageRank является то, что расчет PageRank ведется без учета текстового содержимого документа, а вот структура ссылок web графа задействуется. Таким образом, PageRank позволяет отсортировать все документы в сети по важности еще до получения поискового запроса.

Помимо PageRank на практике реже используют и другие модели ссылочного ранжирования. К ним можно отнести BackRank (модификация PageRank), HITS, HillTop, SALSA. Перечисленные модели задействуют анализ web графа целиком или его части

Современные поисковые системы при реализации комбинируют несколько моделей поиска. Условно можно разделить все модели информационного поиска на две группы. К первой относятся модели, анализирующие текст, а ко второй группе – модели, учитывающие структуру ссылок. Как правило, учет ссылок позволяет оценить авторитетность (важность) ресурса, а текстовый анализ – релевантность запросу.

Список литературы

1. Дударь З. В. Метаконтекстный поиск в internet / З. В. Дударь, В. С. Хапров, А. В. Мусинов // Восточно-Европейский журнал передовых технологий. – 2005. – С. 104–107.

2. Кузнецов М. А. Основные приципы ранжирования Web-ресурсов / М. А. Кузнецов, Т. Т. А. Нгуен // Инновационные технологии в управлении, образовании, промышленности "АСТИНТЕХ-2010". – Астрахань, 2010. – С. 42–44.

3. Ландэ Д. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Снарский, И. В. Безсуднов. – Москва : Книжный дом «ЛИБРОКОМ», 2009. – 264 с.

4. Сегалович И. В. Как работают поисковые системы / И. В. Сегалович. – Режим доступа: http://download.yandex.ru/company/iworld-3.pdf (дата обращения: 30.01.2013), свободный. – Загл. с экрана. – Яз. рус.

References

1. Dudar Z. V., Khaprov V. S., Musinov A. V. Metakontekstnyy poisk v internet [Metacontextual search in internet]. *Vostochno-Yevropeyskiy zhurnal peredovykh tekhnologiy* [East European Journal of Advanced Technologies], 2005, pp. 104–107.

2. Kuznetsov M. A., Nguen T. T. A. Osnovnye pritsipy ranzhirovaniya Web-resursov [Main principles of Web resources ranking]. *Innovatsionnye tekhnologii v upravlenii, obrazovanii, promyshlennosti "ASTINTYeKh-2010"* [Innovative Technologies in Management, Education, Industry "ASTINTEH-2010"]. Astrakhan, 2010, pp. 42–44.

3. Lande D. V., Snarskiy A. A., Bezsudnov I. V. *Internetika: Navigatsiya v slozhnykh setyakh: modeli i algoritmy* [Internet: Navigation in difficult networks: models and algorithms]. Moscow, Book House "LIBROKOM", 2009. 264 p.

4. Segalovich I. V. *Kak rabotayut poiskovye sistemy* [How do search systems work]. Available at: http://download.yandex.ru/company/iworld-3.pdf (accessed 30 January 2013).

УДК 004.032.26 + 338.27

МУЛЬТИАГЕНТНЫЙ МЕТОД УПРАВЛЕНИЯ ЭНЕРГОПОТОКАМИ В ГИБРИДНОЙ ЭНЕРГОСИСТЕМЕ С ИСТОЧНИКАМИ ВОЗОБНОВЛЯЕМОЙ ЭНЕРГИИ

Май Нгок Тханг, аспирант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 65, e-mail: kamaev@unix.cad.vstu.ru

Камаев Валерий Анатольевич, доктор технических наук, профессор, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 65, e-mail: kamaev@unix.cad.vstu.ru