АЛГОРИТМИЧЕСКИЕ ОСНОВЫ РАЗРАБОТКИ ПОИСКОВОЙ СИСТЕМЫ

Трегубов А.А., Кононова Т.С.
Таганрогский Государственный университет Радиотехники,
факультет информационной безопасности, кафедра БИТ, 347928, Россия,
г.Таганрог, ул.Чехова, 2
e-mail: taa_trtu@mail.ru

ALGORITHMIC BASICS OF SEARCH ENGINE DEVELOPMENT

A. A. Tregubov, T. S. Kononova
Department of Information Security, Taganrog State University
of Radio Engineering, ul. Chekhova, 2, Taganrog, 347928, Russia
email: taa_trtu@mail.ru,
phone: (8634) 371905.

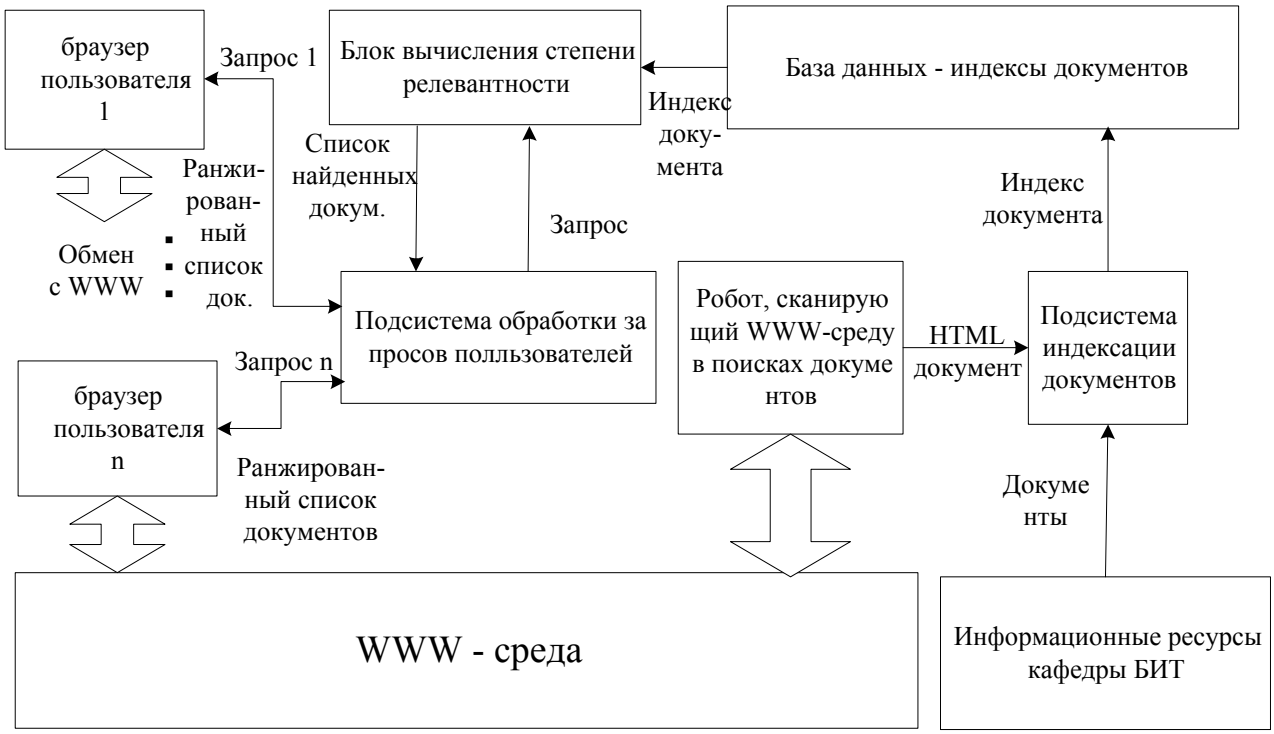
The basics of search engines development are reviewed in this report. A structure of search engine as a part of an electronic library is offered.

Methods of smart search of relevant information based on multi-agent systems and document processing methods are reviewed in the report.

Analysis of major problems of processing, indexing and relevance evaluation is carried out. Statistical indexing, algebraic relevance evaluation and linguistic automaton construction for effective document processing and understanding are considered.

В рамках проекта создания Электронной библиотеки по защите информации на кафедре БИТ разрабатывается поисковая система (ПС). Электронная библиотека будет содержать информацию, относящуюся к области защиты информации, с возможностью дальнейшего ее пополнения и расширения материалами, предоставленными кафедрой БИТ и внешними ресурсами сети Интернет. Ниже представлена структура системы, осуществляющей индексирование документов, поиск релевантной информации. Для работы с внешними ресурсами сети Интернет, представляющей собой постоянно изменяющуюся информационную среду, наиболее перспективными являются методы и алгоритмы интеллектуального поиска релевантной информации. Одним из современных направлений искусственного интеллекта являются мультиагентные системы (МА-системы). Основная составляющая МА-системы – агент, который представляет собой относительно автономную программную компоненту, которая способна самостоятельно определять свои действия, т.е. поведение. Основная идея МА-систем за-

ключается в том, что решаемая МА-системой общая задача декомпозируется на подзадачи составляющих систему агентов. Решение задачи осуществляется как композиция поведения агентов, осуществляющих выбор и реализацию последовательности доступных им действий, направленных на достижение собственных целей. Специальные агенты-роботы будут осуществлять поиск и обработку страниц в сети, извлекать гипертекстовые ссылки на этих страницах, переходить по этим ссылкам, обрабатывать и передавать найденную информацию (HTML-документы).



Структура организации поисковой системы

Работа МА-системы в целом и агентов в частности предполагает наличие механизмов и алгоритмов адаптации агентов, т.е. алгоритмов обучения, при поиске в информационных средах релевантной информации, удовлетворяющей запросу пользователя. Наибольший интерес представляют генетические алгоритмы и эволюционные стратегии, который могут быть использованы для решения общей задачи поиска, управления работой популяции агентов (“создание”, ”размножение”, ”уничтожение” агентов внутри популяции).

В качестве интеллектуального ядра агентов, решающего задачу определения степени релевантности информации (документа), используются нейронные сети прямого распространения — одно и многослойные персептроны. Нейронные сети представляют собой универсальные апроксиматоры функций и эффективное средство распознавания образов. Отличительной чертой нейронных сетей является их способность к обучению на

примерах, без необходимости формулирования четких правил решения задачи. Агенты, наделенные обучаемыми нейронными сетями в качестве механизмов принятия решений, становятся способными к адаптации к окружающей ситуации и интересам пользователя (в данном случае они представлены тематикой искомых документов). Универсальность нейронных сетей позволяет агентам адаптироваться к сложному и заранее неизвестному ландшафту целевой функции, формируемому содержимым информационной среды и запросом пользователя.

При работе агентов во внешней среде, html-документ является базовым элементом. Кроме конечной цели поиска, он выступает также в роли промежуточного этапа, определяя направление пути поиска. Для эффективной работы поисковой системы необходимо выбрать максимальное количество полезной информации из html-документа, для чего надо проводить детальный синтаксический анализ каждого встречающегося документа. В ходе такого анализа в первую очередь обрабатывается список ключевых слов, заданных автором данного документа, текст документа. При установлении соответствия данного документа направлению поиска, проводится дальнейший анализ документа, включающий выборку имеющихся в документе ссылок и оценку этих ссылок путем анализа идущего непосредственно перед ссылкой текста и текста ссылки. Оценка ссылок позволяет определить дальнейший путь поиска, то есть выбрать из всех имеющихся в документе ссылок наиболее точно отвечающих тематике поиска.

При разработке ПС, не зависимо от предполагаемой ее архитектуры, встают две основные проблемы, от эффективности решения которых, кардинально зависит качество создаваемой ПС:

– проблема эффективного семантического анализа текста документа для последующего его индексирования и определения соответствия его запросу пользователя.

– проблема организации эффективного поиска по базе индексов релевантных документов, отвечающих запросу пользователя;

Первая проблема подразумевает разработку метода и алгоритма обработки текстов документов для выделения значимых терминов, определяющих содержание документа, а также определения весовых коэффициентов этих терминов. Данные термины и их веса будут использоваться при создании индекса документа - информации, в сжатом виде представляющей основной смысл документа. Решение второй проблемы сводится к разработке структуры хранения индексов документов, алгоритма поиска по данной базе индексов и алгоритма определения степени релевантности документа.

Поскольку основной лексической единицей текстового документа является слово или термин, существующие методы индексации базируются именно на терминах Основными и наиболее значимыми критериями,

используемыми поисковыми системами для описания индексируемых терминов документа, являются следующие:

- степень присутствия в документе (частота появления),
- специфичность, определяется при уточнении смыслового значения и специфики термина,
- место присутствия в документе (находится в заголовке, подзаголовке, начале документа).

При составлении индекса должны исключаться слова, несущие чисто грамматические функции, общеупотребительные слова, знаки препинания. Общеупотребительные слова встречаются в любом текстовом документе и слабо коррелированны его тематикой. Это такие части речи как предлоги, союзы и т.д.

Для проведения семантического анализа документа во время индексирования и дальнейшего хранения в базе индексов, обрабатываемые термины должны приводиться к специальному виду или канонической форме. [1] В русском языке смысловая информация распределена в словах дискретными «сгустками»: наиболее информационно нагруженными являются обычно начальные морфемы слов (корень), меньше информации несут конечные морфемы (окончание, суффикс), поэтому одним из простейших вариантов приведения слов к специальному виду является отсечение конечной части, а именно: падежных окончаний и суффиксов.

Естественный язык представляет собой нечеткую, сверхсложную семиотическую систему. Поэтому только отсечения окончаний и суффиксов недостаточно, необходима разработка более эффективных алгоритмов представления, обработки и понимания текстов на естественном языке. Наиболее перспективным является построение лингвистического автомата. Идеальный лингвистический автомат должен иметь единую универсальную лингвистическую информационную базу, представляющую собой организованное множество единиц всех лингвистических уровней и способную обслуживать все варианты автоматического анализа и синтеза текста. [2] Основным компонентом лингвистической информационной базы является автоматический словарь, в котором концентрируется основная информация, необходимая для реализации алгоритмов заданного режима. Автоматический словарь представляет собой упорядоченный список лингвистических единиц, каждая из которых снабжена достаточно исчерпывающей с точки зрения задач автоматического анализа и синтеза лексико-грамматической, семантической и семантико-синтаксической информацией. Эта информация формализована в виде кодов. При необходимости каждая лингвистическая единица может быть снабжена стилистической, прагматической, ремо-тематической и прочей информацией, заданной в виде соответствующих кодов.

В зависимости от строя языка и конкретных задач автоматического анализа и синтеза текста в инженерной лингвистике используются следующие типы автоматических словарей.

1. Автоматический словарь машинных основ, представляющих собой модели элементарных лексических знаков.

2. Автоматический словарь словоформ, представляющих собой машинные модели составных знаков.

3. Автоматический словарь словосочетаний (машинных оборотов), являющихся моделями усложненных знаков.

4. Автоматический словарь усложненных структур (синтаксических фреймов), выступающих в роли машинных моделей синтаксических знаков.

Основным элементом каждого автоматического словаря является словарная статья, содержащая в себе всю информацию, характеризующую данную лингвистическую единицу. Эффективность лингвистической информационной базы зависит от трех взаимообусловленных факторов, характеризующих словарную статью: объема лингвистической информации, которая закладывается в словарную статью; способа ее компоновки в словарной статье; организации самой словарной статьи. В использующихся в настоящее время лингвистических информационных базах применяется три способа организации словарной статьи: словарная статья жестко фиксированной длины; словарная статья с плавающими границами; словарная статья с иерархической структурой.

Для разрабатываемой ПС был реализован автоматический словарь машинных основ, представляющий собой перечень машинных основ слов расположенных в произвольном порядке. Словарная статья имеет следующую структуру: собственно машинная основа, цепочка информации к основе, включающая код машинного склонения и, в общем случае, флексию канонической формы. Автоматический словарь реализован в виде таблицы, в которой каждой словарной статье выделяется отдельная строка.

Для решения задачи нахождения канонической формы слова достаточно выделить во входном слове машинную основу, найти ее в словаре, считать из словаря грамматический код, флексию канонической формы и состыковать соответственные машинную форму и флексию. Для получения исходного состояния слова (число, падеж) проводится анализ флексии с использованием кода машинного склонения.

Реализованный словарь позволяет определять падеж, число, склонение, часть речи слов, что увеличивает эффективность не только частотного, но и семантического анализа и понимания текста документов.

При разработке метода индексации необходимо учитывать проблему общности /специфичности. Общность индексации подразумевает составление индекса, максимально отражающего все аспекты содержания документа. Специфичность, наоборот, подразумевает выделение из доку-

ментов только наиболее важных терминов. Общность и специфичность индексации напрямую связаны с общностью и точностью поиска.

В разрабатываемой ПС используется статистический метод индексации. Предположим, что имеется коллекция, состоящая из N документов. Определим функцию tf_{ij} как относительную частоту появления термина t_i в документе d_j.:

$$\text{tfij} = \frac{nt}{n},$$

где nt – число встречаемости термина в документе,
n – число всех терминов в документе.

Выделив множество часто встречающихся терминов, можно построить простейший индекс, содержащий значения функции tf_{ij} для каждого термина в документе. Такой метод индексации ориентирован на максимальную общность поиска, точность поиска при этом будет низкая. Усовершенствование этого метода можно произвести, введя веса терминов, характеризующие их специфичность.

Определим величину df_i как количество документов в коллекции, содержащих термин t_i. Тогда, величина

$$\log\left(\frac{N}{df_i}\right),$$

именуемая инверсной частотой появления термина в документах, может служить величиной, характеризующей специфичность термина t_i (чем меньшая доля документов содержит термин, тем больше ценность t_i как термина, дискриминирующего документы определенного класса). Широко применяемый комбинированный метод индексации [3] определяет веса терминов как величину w_{ij}:

$$w_{ij} = tf_{ij} \cdot \log\left(\frac{N}{df_i}\right).$$

В зарубежной литературе такой метод обозначается как $tf \cdot idf$ метод. В соответствии с ним, наибольший вес имеют термины, которые встречаются достаточно часто, но при этом, сосредоточенные в небольшой доле документов коллекции.

Определение степени релевантности обрабатываемого документа сводится к определению степени соответствия запроса пользователя к текущему индексу. Для этого необходимо запрос пользователя привести к тому же виду что и индекс, т.е. для терминов запроса необходимо применить операцию приведения к канонической форме.

Для определения степени релевантности используется алгебраический метод, который основан на предположении, согласно которому множество документов коллекции может быть представлено набором векторов в векторном пространстве индексируемых терминов. При этом необходимо проводить нормализацию векторов, которая несет дополнительную положительную роль — она позволяет нормализовать документ по его размеру. Без ее проведения документы, имеющие большой объем и, за счет этого, большие степени присутствия терминов в документе, получат преимущество перед малыми документами, реально имеющих большую плотность и состав полезных терминов в своем содержимом.

Запрос пользователя так же представляется в виде вектора в том же пространстве. Степень релевантности документа, т.е. его “похожесть” на вектор-запрос, вычисляется как некоторая мера расстояния между векторами запроса и документа. Одним из удобных вариантов такой меры может служить расстояние Хемминга: $d(\overline{x},\overline{C}) = |x_1 - c_1| + \ldots + |x_n - c_n|$, где x и C вектора запроса и индекса.

Максимальное значение достигается тогда, когда вектора запроса и индекса документа сонаправлены, т.е. содержат одинаковый долевой состав терминов. В ПС применен стандартный подход, когда в качестве весов терминов индекса используются частоты присутствия. Для определения весов терминов в векторе-запросе использована идея комбинированного метода $tf \cdot idf$, при этом веса отсутствующих в запросе терминов полагаются равными 0, а веса терминов, стоящих в запросе с инверсией, берутся со знаком «-».

Другим вариантом определения весов является их подстройка на основе обратной связи с пользователем. В этом случае, пользователю предлагается оценить некоторое подмножество тестовых документов, и, на основе сделанных оценок, провести коррекцию весов с целью уменьшения невязки между автоматической и пользовательской оценками релевантности документов выборки. Для коррекции, например, может быть использован алгоритм обучения однослойного персептрона.

Очень часто запросы пользователей представляют собой сложные конструкции. Для упрощения сложных запросов, включающих в себя скобки, дизъюнкции, конъюнкции, инверсии используется метод обработки, в результате которого сложный запрос разбивается на более простые, содержащие только конъюнкции терминов. Тогда простейшим вариантом определения степени релевантности документа является вычисление расстояний между простыми запросами и индексом документа, из которых выбирается минимальное.

С точки зрения общей эффективности поисковых систем, важной является результирующая точность оценок релевантности содержимого документов. С эффективностью также связан и порог релевантности, определяющий соответствует ли данный документ запросу пользователя или

нет. При задании порога необходимо руководствоваться такими показателями эффективности как общность и точность поиска.

Различные поисковые системы используют различные алгоритмы ранжирования. При определении степени релевантности документов, в данной ПС используются следующие показатели:

- количество слов запроса в индексе документа;
- относительные частоты слов, найденных в индексе;
- удельный вес слов, найденных в индексе.

Для индексирования документов из внешних ресурсов возможно хранение дополнительной информации о самом HTML-документе, индексируемых терминах, например:

- тэги, в которых эти слова располагаются.
- время - как долго страница находится в базе поискового сервера. Много существует в Интернете сайтов, которые “живут” непродолжительное время. Если же сайт существует довольно долго, это означает, что владелец весьма опытен в данной теме и пользователю больше подойдет этот сайт.
- индекс цитируемости - как много ссылок на данную страницу ведет с других страниц, зарегистрированных в базе поисковика.

Работы, ведущиеся в рамках проекта, поддержаны грантом РФФИ № 02-07-90220.

Литература

1. Пиотровский Р.Г., Билан В.Н., Боркун М.Н., Бобков А.К. Методы автоматического анализа и синтеза текста. – Минск: Выш. шк., 1985.– 224 с.
2. Пиотровский Р.Г. Текст, машина, человек. – СПб.: Наука, 1975.– 328 с.
3. Frakes W.B., Baeza-Yates R. Information Retrieval: Data Structures and Algorithms. — EngleWood Cliffs, N.J.: Prentice Hall, 1992.