УДК 004.93

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ В ПРОЦЕССЕ ИНДЕКСИРОВАНИЯ И РАНЖИРОВАНИЯ ТЕКСТА

Д. С. Черезов, Н. А. Тюкачев

Воронежский государственный университет

В статье рассмотрен алгоритм индексирования текста с применением нейронных сетей с целью повышения уровня релевантности документов на этапе ранжирования документов. Нейронная сеть определяет вероятность принадлежности документа одной из тематик, с учетом которой производиться корректировка результата ранжирования.

ВВЕДЕНИЕ

На сегодняшний день, основным методом определения релевантности документа является подсчет и сравнение каких либо статистических данных текста и поискового запроса, таких как количество искомых слов и их месторасположение в документе.

Вследствие чего для поиска в разных секторах интернета создаются специальные сервисы на базе существующей поисковой системы с учетом, каких либо специфик. Так, например, существуют сервисы по поиску среди учебных заведений, по хранилищам исходных программных кодов.

Одним из недостатков существующих систем является отсутствие такого понятия как тематика документа. Внедрение данного определение способствует качественному улучшению результатов ранжирования документов.

1. ИНДЕКСИРОВАНИЕ ДОКУМЕНТА

Реализованная система предоставляет возможность индексирования текстовых документов в формате html. Под индексом документа понимается статистическая информация о документе, массив индексов используемых тегов и слов. Процесс получение индекса документа делится на следующие этапы:

1. Разделение документа на составляющие элементы (теги, параметры, содержание);
2. Получение индексов слов;
3. Подсчет статистики;
4. Выделение ключевых слов документа;
5. Обработка ключевых слов нейронной сетью с целью классификации документа.

2. РАЗДЕЛЕНИЕ ДОКУМЕНТА НА СОСТАВЛЯЮЩИЕ ЭЛЕМЕНТЫ

Элементом документа является: наименование тега, наименование и значение параметра тега, все слова видимые пользователю.

Анализатор реализован в качестве двух конечных автоматов. На вход первому конечному автомату подается весь текст документа. Матрица, являющаяся основой, содержит в себе соответствие между возможными используемыми символами и выделенными типами символов. Помимо этого возникает возможность быстрого обращения к элементам документа.

Второй конечный автомат определяет порядок следования типов слов. Таким образом, получается структура документа. Это является базой для получения статистики документа по ошибкам в структуре документа.

Ошибки в названиях тегов и параметров находятся путем их проверки на существование в спецификации языка HTML. Информация о полученных ошибках так же сохраняется в виде статистики.

Параллельно процессу нахождение ошибок в именах служебных слов происходит определение формата содержимого документа видимого пользователю, в частности тип и видимость.

Процесс разделение документа позволяет ускорить следующий этап — индексацию, так как получение значения слова происходит путем обращения к нему через уникальный идентификатор.

3. ПОЛУЧЕНИЕ ИНДЕКСА СЛОВ

Индекс слова определяет:

1. Идентификаторы возможных корней;

2. Набор возможных окончаний для каждого корня;

3. Часть речи, в которой употребляется слово в документе.

Для получения индекса документа используется конечный автомат, состояния которого представлены в виде графа. Можно выделить два вида состояний: промежуточные и конечные. Промежуточные состояния хранят информацию о переходах системы в другие состояния, конечные состояния хранят информацию о соответствующем индексе слова.

Конечный автомат перебирает все буквы, из которых состоит слово. Состояние автомата описывает текущую букву из слова. Переходы описывают следующую полученную букву при анализе слова. Если конечный автомат заканчивает работу на промежуточном состоянии, то считается, что индекс не найден, иначе возвращается информация об индексе слова, хранящаяся в состоянии.

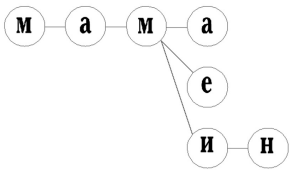

Рис. Визуализация состояний конечного автомата

Состояния конечного автомата хранят все возможные корни слов, количество которых составляет порядка 173 тысяч и так же все возможные окончания и суффиксы для каждого корня. Это является причиной высоких ресурсных затрат на хранение и времени на формирование подобной структуры, однако позволяет получать индексы слов в минимальные промежутки времени и возможность быстро добавлять новые слова.

4. ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ

Ключевыми словами называются слова, которые наиболее полно отображают тематику информации содержащейся в документе. Основным критерием отбора ключевых слов является частота повторений слова, дополнительным — местоположение. Количество ключевых слов вычисляется для каждого документа отдельно по формуле 1.

$$k = \mathrm{trunk}\left(\log\left(\frac{n}{3}\right) + \log(n)\right), \qquad (1)$$

где n — количество слов в документе, k — количество ключевых слов в документе.

Выделение ключевых слов происходит путем подсчета повторов слов и сортировка результатов по убыванию. Выбирается k-тое количество слов, при этом количество их повторений быть удовлетворять неравенство 2. В противном случае документ помечается как малоинформативный, что в свою очередь будет влиять на результаты ранжирования.

$$c \geq k - s, \qquad (2)$$

где c — частота повторения слова, k — количество ключевых слов, s — этап получения ключевых слов.

На основе ключевых слов строиться семантическая сеть. Для этого из текста удаляются полученные на предшествующем этапе ключевые слова и измененный текст опять. Таким образом, операция с получением ключевых слов повторяется три раза. Между всеми словами более высокого и низкого уровня высчитываются веса, представляющие собой количество слов между двумя ближайшими словами.

При обработке документов, для которых количество ключевых слов меньше трех, находятся ключевые слова только первого и второго порядка. Если же количество ключевых слов меньше двух, ищутся ключевые слова только первого порядка.

5. ИСПОЛЬЗОВАНИЕ НЕЙРОННОЙ СЕТИ

Используется трехслойная нейронная сеть с логнормальной функцией отклика. Количество нейронов первого слоя равно тридцати, а количество нейронов выходного слоя определятся в зависимости от области применения, и описывает выделенные тематики документов на этапе обучения сети.

Каждый вход нейрона описывает одно ключевое слово: идентификатор корня, идентификатор окончания и весовые значения на ключевые слова нижнего уровня. Если количество слов, в полученной семантической сети меньше, нежели может вместить формируемая матрица, то оставшиеся строки дополняются нулями.

Полученная в ходе преобразования семантической сети матрица передается на вход нейронной сети. Выходной массив представляет

собой вероятность принадлежности документа к той или иной тематике.

Полученные результаты дополняют статистику документа, и после чего индекс документа записывается в базу данных.

6. РАНЖИРОВАНИЕ

Процесс ранжирования представляет собой сортировку документов содержащие слова из поискового запроса с целью получения наиболее релевантных документов. Критериями релевантности являются:

1. Количество повторений искомых слов;
2. Их месторасположение;
3. Количество ссылок на документ;
4. Количество ссылок с документа;
5. Тематика поискового запроса.

Для удобства сортировки все критерии имеют свое весовое значение, которые в ходе работы суммируются.

6. ПРОВЕРКА ПОИСКОВОГО ЗАПРОСА

При получении поискового запроса происходит получение индексов всех входящих в него слов. Так как порядок получения документов из базы данных содержащих ключевые слова существенно определяет время работы процесса, то используется сортировка слов запроса, с целью выявления наиболее редко встречающихся слов. Для этого используются следующие критерии:

- Часть речи;
- Длина слова;
- Количество однокоренных слов.

Слова, для которых не найдены индексы или же которые используются в каждом документе, удаляются из запроса.

Базовым алгоритмом анализа текста является полнотекстовый поиск искомых слов и построение графа связей документов для определения ранга документа в сети Интернет.

Полнотекстовый анализ текста определяет отношение количества найденных поисковых слов к объему документа с учетом месторасположения слов. Учитываемые места расположения слов:

- Заголовок документа;
- Заголовок текста;
- Текст документа.

При полнотекстовом поиске учитывается статистика документа. Наличие ошибок в словах, в структуре документа, в служебных словах ведет к уменьшению значения оценки документа.

Для определения ранга документа в сети интернет строиться ориентированный взвешенный граф, вершины которого описывают документы, а переходы представляют ссылки с одного документа на другой.

Веса переходов определяются в два этапа. На первом этапе вес перехода определяется на основании статистических данных. На втором этапе происходит корректировка весов для всех вершин поочередно. Так у вершины имеющее входящие переходы с большими весовыми значениями, то веса исходящих переходов так же повышаются.

Граф строиться для каждого документа отдельно и для уменьшения времени работы при повторном анализе данного документа, полученные результаты сохраняются.

7. АНАЛИЗ ЗАПРОСОВ С ПРИМЕНЕНИЕМ НЕЙРОННОЙ СЕТИ

К поисковым запросам выдвигается требование — наличие трех и более слов в поисковом запросе. Однако это требование не является критическим.

Нейронная сеть, как и на этапе классификации документа, состоит из трех слоев и использует логнормальную функцию отклика. Но основным отличием является формат подаваемых данных и как следствие отдельный процесс обучения сети. Так матрица отражает только лишь используемые состояния поисковых слов. Результатом работы матрицы является вероятность принадлежности поискового запроса тому или иному запросу.

Если запрос состоит из трех слов, то их индексы подаются на вход нейронной сети в идее матрицы. Каждая строка описывает одно слово. Количество строк в матрице равно пятнадцати, если количество слов больше количества строк, то в матрице ни как не отображаются. При этом используется отсортированный список поисковых слов, полученный на этапе разбора запроса. В случае, когда количество поисковых слов меньше пятнадцати, то матрица дополняется нулевыми значениями.

Подача на вход нейронной сети матрицы содержащей только одно — два слова является малоэффективным. Поэтому, в случае, когда запрос состоит из одного или двух слов произ-

водиться поиск наиболее близких слов способных дополнить запрос.

Рассматриваются семантические сети, в которых используются искомые слова и выделяются наиболее часто встречаемые слова таким образом, что бы полностью дополнить ими матрицу описывающую запрос.

Документы, используемые в процессе дополнения запроса, берутся случайным образом из всего списка релевантных документов отсортировано по результатам полнотекстового поиска и поиска ранга документа.

Полученные результаты для анализируемого запроса сохраняются в базе данных.

Завершающим этапом ранжирования является корректировка ранее полученных результатов с учетом темы запроса и тем документов. Таким образом, ответ ранжирования представляется в виде блоков документов по двадцать документов в каждом.

Каждый блок содержит определенное количество документов каждой темы. Количество документов в блоке относящихся одной тематике определяется вероятностью принадлежности запроса к этой тематике.

Сортировка документов в блоках происходит на основании результатов полнотекстового поиска и поиска ранга документов.

ЛИТЕРАТУРА

1. Автоматическая обработка текстов //http://www.aot.ru/

2. *Demiriz A., Bennett K.P., Embrechts M.J.* (1999). Semi-supervised clustering using genetic algorithms. In Artificial Neural Networks in Engineering (ANNIE-99), P. 809—814.

3. *Sugato B.* Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments http://www.cs.utexas.edu/users/sugato/papers/sugato-phdthesis.pdf

4. *Косинов Д., Апрелов Б.* Применение методов отбора признаков при решении задачи классификации документов // Материалы II конференции “Современные проблемы прикладной математики и математического моделирования”. 2007.

Статья принята к опубликованию
25 октября 2007 г.