Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК_____

Белоусов
О.А.

Профилирование пользователя Интернет

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 02 Системный анализ, управление и обработка информации (по отраслям)

_____

Научный руководитель
Герман О.В._____
кандидат технических наук, доцент

_____

Минск 2014

Нормоконтроль

_____

_____

ВВЕДЕНИЕ

Ежедневно пользователи всемирной паутины получают массу информации. Иногда она бывает полезной, однако большая ее часть не представляет интереса для человека, который ее просматривает. Это связано с тем, что каждый человек – это личность, он уникален. То, что представляется занимательным одному, другому может показаться бессмысленным. У каждой статьи найдется благодарный читатель, однако вероятность того, что им будет именно тот, кому она в данный момент предъявлена, очень невелика.

Профилирование пользователя – это разумное ограничение предъявляемой посетителю информации с целью выделения более важного для него содержания.

Это достаточно известная и распространенная задача, решаемая в настоящее время различными способами. Её суть заключается в том, что пользователю произвольного информационного Интернет-ресурса предоставляют не весь контент, а в первую очередь то, в чем, предположительно, он может быть заинтересован. Предположение обычно строится на основе многих факторов: документов, которые пользователь смотрел в прошлом, его географического положения, приватной информации из личного профиля пользователя и т.д.

Задачей профилирования является правильный отбор пар «пользователь – набор отображаемых данных» путем отсеивания неинтересной пользователю информации.

Решение этой задачи позволит потребителям услуг тратить меньше времени на просмотр информации и больше – на ее практическое применение.

В данной диссертациипредложен новый метод профилирования пользователя сети Интернет с помощью алгоритма отбора релевантных документов, основанный на использовании корреляционной матрицы, характеризующей связи между ключевыми словами предметной области, к которой относится документ.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Классификация документов — одна из задач информационного поиска, заключающаяся в отнесении документа к одной из нескольких категорий на основании содержания документа. Она может осуществляться полностью вручную, либо автоматически с помощью созданного вручную набора правил, либо автоматически с применением методов машинного обучения.

Все существующие методы классификации базируются на понятии “расстояния”.Традиционно в качестве модели документа использовалось линейное векторное пространство. Каждый документ представлялся как вектор, то есть массив всех и наиболее часто встречающихся слов, а затем для расчета использовалось, например, евклидово расстояниеили мера Жаккара.

Однако эти расстоянии сложны для измерения. Они не учитывают связи между словами, что является крайне важным для эффективного решения задачи профилирования пользователя сети Интернет.

Поэтому в своей работе я использую новый подход, который учитывает корреляцию (связь) между слова, что позволяет существенно больших результатов.

Представленный мной алгоритм отбора релевантных документов, основан на использовании корреляционной матрицы, характеризующей связи между ключевыми словами предметной области, к которой относится документ. Отыскивается сингулярное разложение корреляционной матрицы, позволяющее найти вектор диагональной матрицы (называемый вектором главных компонент), уникально представляющий данный документ. Все документы с похожими векторами образуют семантически общий кластер. Задача состоит в том, чтобы отобрать требуемый кластер, соответствующий поисковому запросу пользователя. Эта задача сводится к оценке близости двух векторов и для ее решения можно использовать известные метрики, в том числе и нечеткие типа Сугено или Мамдани. Таким образом, достигается решение задачи профилирования.

Результаты, достигаемые при реализации данного подхода, позволят повысить эффективность поиска и выдачи информации для каждого пользователя сети Интернет

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Решение нашей задачи мы должны начать с построения матрицы ключевых слов, у которой столбцы – ключевые для нас слова (Kn), а строки – номера абзацев анализируемого нами документа:

	K1	K2	K3	…	Kn
1	1	0	1		
2	0	1	0		
3	1	0	0		
…					
m					

Таблица 1 – Матрица ключевых слов

При наличии того или иного ключевого слова в конкретном абзаце на их пересечении будет поставлена либо 1 – слово присутствует, либо 0 – слово отсутствует

Далее нам следует провести корреляционный анализ, который состоит в определении степени связи между двумя случайными величинами X и Y. В качестве меры такой связи используется коэффициент корреляции. Коэффициент корреляции оценивается по выборке объема n связанных пар наблюдений (xi, yi) из совместной генеральной совокупности X и Y. Существует несколько типов коэффициентов корреляции, применение которых зависит от измерения (способа шкалирования) величин X и Y

С помощью MicrosoftExcelмы можем произвести корреляционный анализ и найти парные линейные коэффициенты корреляции

Для этого используется специальная функция КОРРЕЛ (массив1; массив2),

где массив1 – ссылка на диапазон ячеек первой выборки (X);

массив2 – ссылка на диапазон ячеек второй выборки (Y).

Затем для решения нашей задачи отбора релевантных документов мы должны отыскать сингулярное разложение нашей корреляционной матрицы,

позволяющее найти вектор диагональной матрицы (называемый вектором главных компонент), уникально представляющий данный документ. Все документы с похожими векторами образуют семантически общий кластер.

Сингулярное разложение можно произвести, например, с помощью онлайн сервиса http://www.dotnumerics.com/MatrixCalculator/default.aspx

Итак, пусть имеются документы по определенной тематике. У каждого документа будет свое сингулярное разложение, но главные компоненты будут похожи. Значит, кластер должны образовывать документы с похожими главными компонентами. Далее допустим, что у нас есть несколько кластеров. Тогда можно построить нейросеть для распознавания каждого кластера.

Для построения нейросетей, с помощью которых будет происходить распознавание каждого отдельного кластерами (включающего документы с похожими главными компонентами), мы воспользуемся алгоритмом обратного распространения ошибки.

Алгоритм обратного распространения ошибки является одним из методов обучения многослойных нейронных сетей прямого распространения, называемых также многослойными персептронами. Многослойные персептроны успешно применяются для решения многих сложных задач.

Обучение алгоритмом обратного распространения ошибки предполагает два прохода по всем слоям сети: прямого и обратного. При прямом проходе входной вектор подается на входной слой нейронной сети, после чего распространяется по сети от слоя к слою. В результате генерируется набор выходных сигналов, который и является фактической реакцией сети на данный входной образ. Во время прямого прохода все синаптические веса сети фиксированы. Во время обратного прохода все синаптические веса настраиваются в соответствии с правилом коррекции ошибок, а именно: фактический выход сети вычитается из желаемого, в результате чего формируется сигнал ошибки. Этот сигнал впоследствии распространяется по сети в направлении, обратном направлению синаптических связей. Отсюда и название – алгоритм обратного распространения ошибки. Синаптические веса настраиваются с целью максимального приближения выходного сигнала сети к желаемому.

Обучение методом обратного распространения ошибки происходит в соответствии со следующими пунктами:

1. Подать на входы НС образец выбранный случайным образом и в режиме обычного функционирования НС рассчитать ее выходы.

2. Рассчитать дельту ошибки для выходного слоя по формуле:

$$\delta_i^{(N)} = (y_i^{(N)} - d_i) \cdot \frac{dy_i}{ds_i} \quad (1)$$

3. Рассчитать изменения весов для выходного слоя N по формуле:

$$\Delta w_{ij}^{(n)} = -\eta \cdot \delta_j^{(n)} \cdot y_i^{(n-1)} \quad (2)$$

4. Рассчитать ошибки и изменения весов для всех остальных слоев по формулам:

$$\delta_j^{(n)} = \left[\sum_k \delta_k^{(n+1)} \cdot w_{jk}^{(n+1)} \right] \cdot \frac{dy_j}{ds_j} \quad (3)$$

5. Скорректировать все веса нейронной сети по формуле:

$$w_{ij}^{(n)}(t) = w_{ij}^{(n)}(t-1) + \Delta w_{ij}^{(n)}(t) \quad (4)$$

6. Если ошибка сети существенна, то перейти на шаг 1. В противном случае – конец обучения

Изменение весов нейронной сети можно вести также с помощью формулы:

$$\Delta w_{ij}^{(n)}(t) = -\eta \cdot (\mu \cdot \Delta w_{ij}^{(n)}(t-1) + (1-\mu) \cdot \delta_j^{(n)} \cdot y_i^{(n-1)}) \quad (5)$$

где μ - коэффициент инерционности. Эта формула позволяет «гладко» изменять веса, что приводит к понижению перепадов ошибки. Её целесообразно применять в конце обучения, когда ошибка близка к заданной.

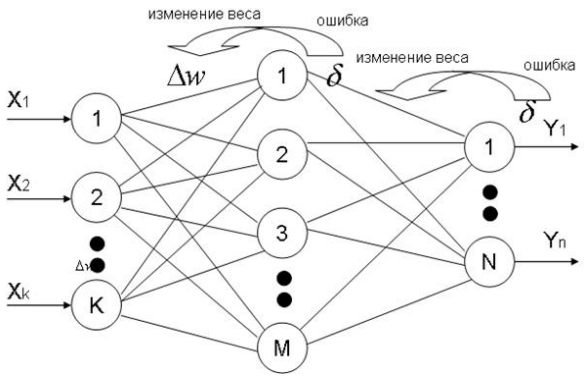

Рисунок 1 – Пояснение алгоритма обратного распространения ошибки

Таким образом, с помощью данного алгоритма можно успешно анализировать и распознавать каждый кластер документов с целью выявления степени его соответствия той или иной тематике.

Последующее создание, обучения и моделирование нейронной сети согласно данному алгоритмуможет быть осуществлено в среде MatLAB.

Построение нейронной сети начинается с создания модели сети. Для создания модели сети в системе MatLAB применяется функция newff.

Для изменения структуры нейронной сети существует ряд команд. Например: net.layers{1}.size=12; и net.inputs{1}.size=36 и т.д.. Здесь net – созданная сеть; layers{1}.size – число нейронов в первом слое; inputs{1}.size – число элементов вектора входа.

Для обучения сети используется функция train, а также различные алгоритмы обучения. При обучении необходимо настроить параметры обучения: epochs – максимальное количество циклов обучения; goal – предельное значение критерия обучения; lr – параметры скорости настройки; max_fail – максимально допустимый уровень превышение ошибки контрольного подмножества по сравнению с обучающим; min_grad – минимальное значение градиента; show – интервал вывода информации измеренный в циклах; time – предельное время обучения; mem_reduc – позволяет экономить объем используемой памяти; mu – начальное значения для коэффициента μ.

Моделирование сети после её обучения осуществляется с помощью функции sim(net, t), где net – созданная нейронная сеть; t – данные для обучения.

ЗАКЛЮЧЕНИЕ

Результатом поэтапной реализации данного проекта стал алгоритм отбора релевантных документов, позволяющий произвести правильный отбор пар «пользователь – набор отображаемых данных» путем отсеивания неинтересной или ненужной пользователю информации.

Использование данного метода профилирования позволит пользователямсети Интернет тратить меньше времени на просмотр информации и больше на ее практическое применение.

В процессе изучения предметной области были рассмотрены уже существующие принципы отбора ключевых слов, алгоритмы проведения их сходства, изучен аппарат нейронныхсетей.

Результаты, достигаемые при реализации данного подхода, позволят повыситьэффективность поиска и выдачи информации для каждого пользователя сети Интернет. Данный алгоритм позволит повысить производительность средств распространения рекламных информационных материалов в Интернет и эффективность рекламного и информационного воздействия на пользователей, может быть использован в части удовлетворения поисковой активности пользователя, для создания интеллектуального «электронного учителя».

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1] Белоусов О.А., Герман О.В. Задача профилирования пользователя сети интернет в контексте системы дистанционного обучения: Тезисы докл. к VII международной научно-методической конференции «Высшее техническое образование: проблемы и пути развития»– Минск, 2014 – С.122 – 123.