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

УДК 004.91

Потараев
Виктор Витальевич

Модели, методы и программное средство обработки текстовой информации на основе семантического анализа

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Серебряная Л.В.
к.т.н., доцент

Минск 2015

КРАТКОЕ ВВЕДЕНИЕ

На протяжении последних двадцати лет количество текстовых документов, представленных в цифровом виде, очень резко возросло. Классификация является задачей, которую довольно часто приходится решать при работе текстом. Основной задачей текстовой классификации является разделение неструктурированного множества документов на группы в соответствии с содержимым документов. Классификация текстов имеет множество применений, например, разделение электронных сообщений по категориям, фильтрация спама определение темы текста и др.

Наиболее простым для понимания и реализации методом классификации текстовой информации является метод частотного анализа, который опирается на частоту встречаемости различных слов в тексте. Чуть более сложными методами логично считать модификацию частотного анализа, например, частотно-контекстный анализ.

Методы, основанные на семантическом анализе, ещё сложнее. Семантика – это часть лингвистики, исследующая функции слов, их связь между собой и окружающей действительностью. Прослеживаются эти механизмы с помощью семантического анализа.

Одним из методов классификации, который анализирует связи между понятиями, является метод, основанный на применении нейронных сетей [6]. Нейронные сети – это одно из направлений исследований в области искусственного интеллекта, основанное на попытках воспроизвести нервную систему человека. На теорию нейронных сетей существенно повлияла модель персептрона, для обучения которой есть несколько способов, одним из который является метод обучения с обратным распространением ошибки, и модель нейронной сети Хопфилда.

Ещё одним из способов обработки текстовой информации является обработка при помощи семантических сетей. Однозначное определение семантической сети в настоящее время отсутствует. В инженерии знаний под ней подразумевается граф, отображающий смысл целостного образа. Узлы графа соответствуют понятиям и объектам, а дуги – отношениям между объектами.

Диссертационная работа посвящена разработке алгоритмов и программного средства, позволяющих решать задачи классификации текстовой информации как можно более точно.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является повышение точности классификации текстовой информации на базе персональных компьютеров общего назначения.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Исследовать существующие модели и методы построения и использования нейронных сетей.
2. Исследовать существующие модели и методы построения и использования семантических сетей.
3. Разработать методы классификации текстов на основе нейронных сетей.
4. Разработать метод классификации текстов на основе семантической сети.
5. Реализовать программное средство, позволяющее решать задачу классификации текстовой информации.

Объектом исследования является классификация текстовой информации.

Предметом исследования являются модели и методы анализа текстовой информации, а также алгоритмы её классификации.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования нейронных и семантических сетей для проведения семантического анализа текстовой информации и решения задачи классификации. Нейронные сети позволяют представить смысл текста в виде сложных математических формул. Семантические сети представляют смысл текста в наглядной и формализованной форме.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии научно-техническими заданиями и планами работ кафедры «Программное обеспечение информационных технологий», и хозяйственными договорами с предприятиями Республики Беларусь:

1. Проведение исследований по теме «Модели, методы и программное средство обработки текстовой информации на основе семантического анализа» (ГБ № 11-2004, № ГР 20111065, научный руководитель НИР – В. В. Бахтизин).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Л. В. Серебряной заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Работа выполнялась в соответствии с научно-техническим заданием и планом работ кафедры «Программное обеспечение информационных технологий» по теме «Модели, методы и программное средство обработки текстовой информации на основе семантического анализа» (ГБ № 11-2004, № ГР 20111065, научный руководитель НИР – В. В. Бахтизин).

Опубликованность результатов диссертации

По теме диссертации опубликовано 3 печатных работ, из них 1 работа в сборнике материалов международной конференции, 1 работа в сборнике материалов конференции БГУИР, 1 статья в рецензируемом издании.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников и списка публикаций автора. В первой главе представлен анализ предметной области, выявлены основные существующие модели и методы анализа текстовой информации. Во второй главе предложены методы анализа и классификации текстов. Третья глава посвящена разработке архитектуры и алгоритмов для программного средства по теме исследования. В четвертой главе предложена практическая реализация ПС для применения и сравнения методов классификации текстовой информации.

Общий объем работы составляет 75 страниц, из которых основного текста – 50 страниц, 29 рисунков на 20 страницах, 1 таблица на 2 страницах и список использованных источников из 30 наименований на 3 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** описаны модели представления текстовой информации, приведены их преимущества и недостатки. Отдельно рассмотрены история развития и некоторые модели нейронных сетей и семантических сетей.

Одним из ключевых понятий, характеризующим выбор того или иного метода анализа текстовой информации, а также реализацию конкретного варианта поиска информации, является модель поиска. Можно выделить следующие наиболее популярные модели поиска: булевские, векторные, вероятностные модели и сети вывода.

Нейронные сети имеют довольно длинную историю развития. В 1949 году канадский физиолог и психолог Хебб высказал идеи о характере соединения нейронов мозга и их взаимодействии. Он первым предположил, что обучение заключается в первую очередь в изменениях силы синаптических связей. В 1954 году в Массачусетском технологическом институте с использованием компьютеров Фарли и Кларк разработали имитацию сети Хебба. В 1957 году Розенблаттом были разработаны математическая и компьютерная модели восприятия информации мозгом на основе двухслойной обучающейся нейронной сети. В 1975 году Вербос разработал метод обратного распространения ошибки, который позволил эффективно решать задачу обучения многослойных сетей и впервые реализовать «сложение по модулю 2» при помощи нейронной сети. В 1982 году в сети Хопфилда удалось достичь двусторонней передачи информации между нейронами.

Начиная с конца 1950-ых гг. были созданы и применены на практике десятки вариантов семантических сетей. Несмотря на то, что терминология и их структура различаются, существуют сходства, присущие практически всем семантическим сетям, например:

– узлы семантических сетей представляют собой концепты предметов, событий, состояний;

– различные узлы одного концепта относятся к различным значениям, если они не помечено, что они относятся к одному концепту;

– дуги семантических сетей создают отношения между узлами–концептами (пометки над дугами указывают на тип отношения).

В конце 1970-х семантические сети получили широкое распространение.

Семантический анализ заключается в определении информативности текстовой информации и выделении информационно-логической основы текста. Представление логической основы текста при помощи семантической сети более наглядно, чем при помощи нейронной сети.

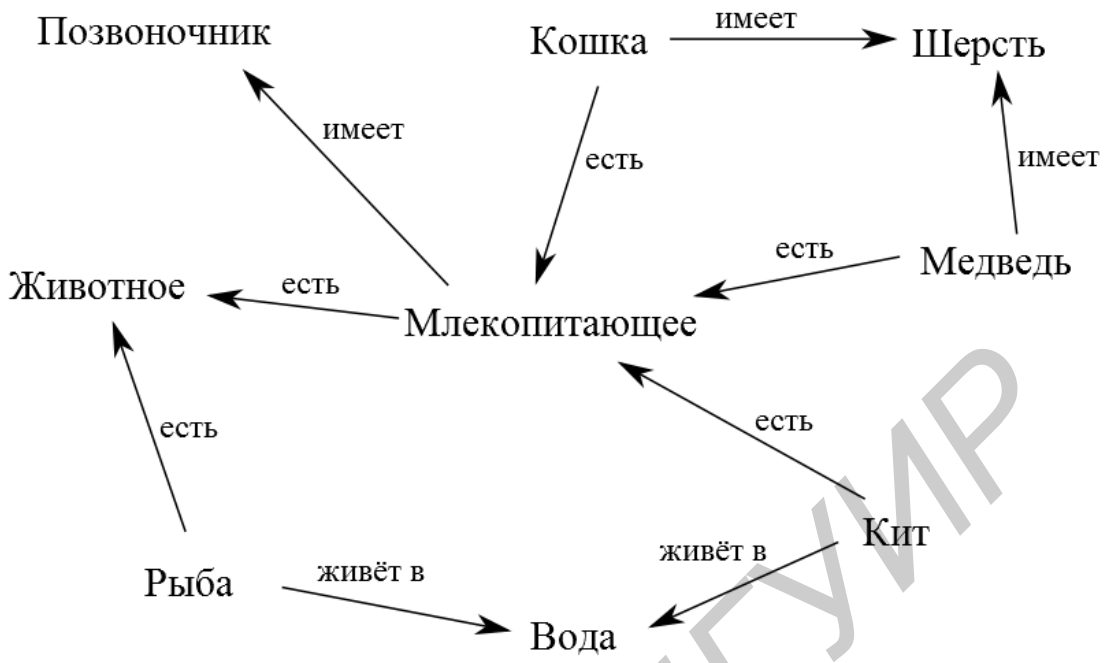

Рисунок 1 – Пример семантической сети

Вторая глава посвящена анализу моделей и методов обработки текстовой информации на основе семантического анализа. В ней подробно рассмотрен персептрон, обучаемый по методу обратного распространения ошибки, нейронная сеть Хопфилда, а также рассмотрены различные модели семантических сетей. Предложено применение данных моделей для решения задачи классификации текстовой информации.

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 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.

Модель Хопфилда (J.J.Hopfield, 1982) занимает особое место в ряду нейросетевых моделей. В ней впервые удалось установить связь между нелинейными динамическими системами и нейронными сетями. Образы памяти сети соответствуют устойчивым предельным точкам (аттракторам) динамической системы. Особенно важной оказалась возможность переноса

математического аппарата теории нелинейных динамических систем (и статистической физики вообще) на нейронные сети.

Семантические сети возникли как попытка визуализации математических формул. Основным представлением для семантической сети является граф. Количество типов отношений в семантической сети определяется её создателем, исходя из конкретных целей. В реальном мире их число стремится к бесконечности. Каждое отношение является, по сути, предикатом, простым или составным. Часто в семантических сетях требуется определить отношения синонимии и антонимии. Эти связи либо дублируются явно в самой сети, либо определяются алгоритмической составляющей. Проблема поиска решения в семантической сети сводится к задаче поиска фрагмента сети, соответствующего подсети, отражающей поставленный запрос. Это, в свою очередь, обуславливает сложность поиска решения в семантических сетях.

В **третьей главе** описана архитектура разрабатываемого программного средства, предложены способы предварительной обработки текста, а также описаны алгоритмы классификации текстовой информации на основе нейронных сетей и на основе семантической сети.

Предложенная предварительная обработка заключается в удалении из текста стоп-слов (не несущих смысловой нагрузки), а также стемминг (замена слов на их основу).

В случае применения нейронной сети с обратным распространением ошибки для решения задачи классификации, выходам нейронных сетей можно сопоставить классы, к которым относятся обучающие тексты. Начальные веса связей задаются равными случайным числам от 0 до 1. После обучения, во время использования сети для классификации текстов, для каждого из M слов текста производится поиск входа, который при обучении соответствовал этому слову. Если вход найден, то на него подаётся входной сигнал 1. После окончания обучения, то есть когда для всех текстов обучающей выборки нейронная сеть выдаёт корректный сигнал на выходе, она готова к использованию. В качестве результата классификации используется группа текстов, соответствующая выходу нейронной сети, который при подаче входного сигнала превышает некоторое пороговое значение T.

В нейронной сети Хопфилда начальные веса связей задаются по формуле, поэтому процесс обучения не является итерационным. После расчёта матрицы весов связей, нейронная сеть готова к выполнению классификации. Классификация представляет собой итерационный процесс, при котором нейронная сеть последовательно меняет свои состояния. Это происходит до тех пор, пока сеть не попадёт в некоторое стационарное состояние, соответствующее одному из обучающих сигналов. Необходимо отметить, что условие прекращения итераций может никогда не выполниться, и поэтому классификацию обычно ограничивают определённым максимальным числом итераций.

В качестве семантической сети, используемой для классификации текстовой информации, можно использовать сеть, узлами которой являются

некоторые понятия, выражаемые подлежащими и дополнениями, а связями – сказуемые. Каждому узлу такой сети соответсвуют не конкретные слова, а концепты, то есть сущности, которые можно выразить некоторым множеством синонимов. Связь-сказуемое выходит из подлежащего и заканчивается на дополнении.

Пример семантической сети для текста, состоящего из предложений «Данная функция хорошо подходит для математических преобразований» и «После упрощения функция принимает следующую форму», представлен на рисунке 2.

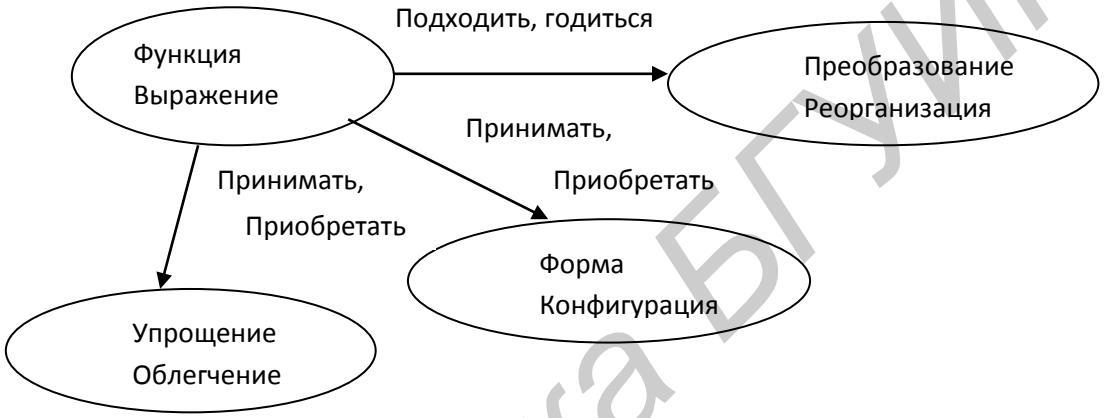

Рисунок 2 – Пример семантической сети для двух предложений

На основании присутствия в сетях, составленных для двух текстов, одной и той же комбинации «узел-связь-узел», можно говорить о некоторой их смысловой близости. В случае отсутствия связи смысловая близость так же может присутствовать, поэтому формула расчёта близости должна учитывать количество совпадающих компонентов предложения.

В **четвертой главе** рассмотрена практическая реализация программного средства классификации текстовой информации, представлены результаты экспериментального исследования разработанных алгоритмов. Система построена по модульно-функциональному принципу.

Исследования разработанных алгоритмов показали, что наиболее точные результаты классификации достигаются при использовании алгоритма, основанного на семантической сети. Этот алгоритм обрабатывает целиком весь текст и является самым медленным. Нейронная сеть Хопфилда является труднонастраиваемой и зачастую выдаёт неверный результат, хотя при небольшом количестве классов результаты довольно хорошие. В целом, использование персептрона, обучаемого по методу с обратным распространением ошибки, является более эффективным методом в сравнении с нейронной сетью Хопфилда, т. к. при сравнимом времени работы алгоритма даёт лучшие результаты, хотя и не такие хорошие, как у алгоритма с использованием семантической сети.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Семантический анализ заключается в определении информативности текстовой информации и выделении информационно-логической основы текста. В данной работе было предложено применение методов семантического анализа текстовой информации для решения задачи классификации.
2. Разработаны алгоритмы классификации текстовой информации на основе нейронных сетей и на основе семантической сети. Эти алгоритмы используют способ обучения с учителем. Показано, что они успешно справляются с поставленной задачей. Алгоритм, использующий семантическую сеть, даёт более высокое количество верных результатов.
3. Разработано программное средство, реализующее разработанные алгоритмы и позволяющее не только получать определённые результаты для различных алгоритмов при различных значениях параметров, но и в удобной форме реализован подбор оптимальных параметров благодаря возможности построения графиков.
4. Проведено сравнение разработанных алгоритмов. Показано, что разработанный метод классификации текстовой информации на основе семантической сети является более точным и более медленным, чем методы, основанные на нейронных сетях.
5. Разработанные алгоритмы имеют свои недостатки, которые можно устранять по мере дальнейших исследований.

Рекомендации по практическому использованию результатов

1. Разработанные методы и алгоритмы классификации текстовой информации могут быть использованы для фильтрации спама, выборки новостей по заданной теме, для классификации научных статей, определения темы текста и при решении других задач.
2. В случае определённой модификации, разработанные алгоритмы могут быть использованы при выборе текстов для изучения.
3. Разработанное программное средство позволяет подбирать оптимальный алгоритм для решения каждой конкретной задачи.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Потараев, В.В. Применение частотно-контекстной классификации текстовой информации при выборе текстов для изучения / В.В. Потараев //

Материалы VIII Международной научно-методической конференции «Дистанционное обучение - образовательная среда XXI века». – Минск: БГУИР, 2013. – с. 340-341.

2-А. Потараев, В.В. Оценка эффективности применения частотно-контекстного анализа текстовой информации / В.В. Потараев // Компьютерные системы и сети: материалы 50-ой научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2014. – с. 33-34.

3-А. Потараев, В.В. Метод классификации текстовой информации на основе семантической сети / В.В. Потараев // Апробация.– 2016. №1 (в печати).