Система классификации текстов NNCS

© Авдейчик В.Г., Чернявский А.Ю., Шмелёв А.С.

«Бинейро»
mail@bineuro.ru

Аннотация

Эта статья является отчётом об участии компании «Бинейро» в семинаре РОМИП'2005. Представлено краткое описание используемой системы, а также результаты её прогона по дорожке классификации веб-сайтов.

1. Введение

Компания «Бинейро» в 2005 году участвовала в семинаре РОМИП впервые. Компанией был представлен макет системы контекстно-зависимой классификации и поиска документов по запросу - "NNCS" (Neural Network Classification & Search). В основе NNCS лежит модель представления текстовых документов в виде семантических векторов, получаемых с помощью специальной рекуррентной нейронной сети. Указанная модель была разработана специалистами компании в 2004 году. В 2005 году на указанную модель компанией получен патент №45579 «Устройство кодирования семантики текстовых документов». Предлагаемый подход может использоваться для различных задач текстового анализа: классификация, кластеризация, поиск.

Главной целью участия в РОМИП была апробация работы нового алгоритма семантического кодирования, а также сравнение качества работы системы с другими системами подобного класса.

2. Краткое описание системы

Работу системы, используемой для классификации текстов, можно разделить на 3 этапа:

- Предварительная обработка текстов
- Получение семантических код-векторов текстов
- Контекстно-зависимая классификация на основе полученных семантических векторов

Предварительная обработка текстов.
На этапе предварительной обработки текстов происходит конвертация html-текста в формат txt, а также морфологический анализ текста (приведение слов к нормальной форме).

Получение семантических код-векторов текстов.
Для получения семантических векторов документов используется рекуррентная нейронная сеть специального вида, являющаяся модификацией известной нейронной сети Хопфилда.

Формирование топологии и связей сети происходит на основе анализа какого-либо корпуса текстов. Множество вершин этой сети находится во взаимнооднозначном соответствии со словарем концептов, сгенерированным в результате морфологического анализа словаря корпуса документов. Веса определяются специальным образом и зависят от статистики совместной встречаемости слов-концептов.

На вход такой сети подаётся нормированный вектор, элементами которого являются статистические веса слов-концептов кодируемого документа. Окончательным код-вектором документа является положение равновесия сети. В силу определённого алгоритма формирования весов, оно всегда существует.

Контекстно-зависимая классификация на основе полученных семантических векторов.
В результате работы нейронной сети каждому документу из обучающей и тестовой выборки ставится в соответствие действительный вектор, размерность которого совпадает с размерностью словаря концептов. Были проведены эксперименты с различными методами классификации с учителем и, как наиболее подходящий к поставленной задаче и экономичный в вычислениях, был выбран следующий:

Каждой категории ставится в соответствие семантический вектор документа, сформированного из всех документов данной категории, записанных подряд. Далее для каждого документа классифицируемой выборки по определённому правилу выбираются пять или меньше категорий, чьи семантические вектора наиболее близки к семантическому вектору документа в смысле стандартной евклидовой метрики

3. Эксперимент для РОМИП и его результаты.

Система NNCS участвовала в дорожке по классификации веб-сайтов. Каждой системе-участнику предоставлялся список категорий, обучающая выборка и множество сайтов из коллекции narod.ru. Было необходимо каждому сайту из коллекции присвоить категорию из этого списка с учётом обучающей выборки.

Один и тот же сайт мог относиться сразу к нескольким категориям, или не относится ни к одной из них. Поэтому ответом являлся упорядоченный список (до 5 категорий) для каждого из классифицируемых сайтов.

Множество категорий сформировано на основе подмножества русскоязычных рубрик каталога DMOZ.

Все программные компоненты системы NNCS, а также вспомогательные компоненты для работы с коллекциями DMOZ и narod.ru были реализованы на языке программирования «C#». В качестве корпуса для формирования нейронной сети были взяты все документы предоставляемой коллекции. Словарь концептов был сформирован из части встречающихся в корпусе прилагательных и существительных в нормальной форме. Его размер составил ~60000 слов. Для всех документов были получены семантические вектора, и для их классификации был использован описанный выше алгоритм. Результаты оценки работы системы представлены в таблице.

	AND	OR
F1 (macro average)	0.182	0.337
Recall	0.451	0.390
Precision (macro average)	0.108	0.270
Error	0.010	0.011
F1	0.191	0.302
Recall (macro average)	0.556	0.448
Accuracy	0.989	0.988
Precision	0.143	0.330

4. Дальнейшие планы

Участие в РОМИП'2005 являлось неоценимым опытом для использования разработанной системы в решении реальных задач текстового анализа. Результаты прогона системы NNCS достаточно интересны относительно общих показателей всех участников. Это открывает перспективы дальнейшей работы по усовершенствованию применяемых системой алгоритмов. Одним из таких направлений является доработка алгоритмов кластеризации и классификации. В планы компании входит оптимизация системы и ее апробация на других дорожках РОМИП, в частности, на дорожках поиска по запросу по Веб-коллекции документов.

System of text classification NNCS

© V.Avdeychik, A.Chernavsky, A.Shmelev

«Bineuro»
mail@bineuro.ru

Abstract

This paper is the report on participation in ROMIP'2005 of "Bineuro". The brief description of our system, and also results of its run on classification of websites is submitted.