Нечеткий поиск на основе нейронной сети Хемминга

Е.С.Борисов (http://mechanoid.narod.ru/)

В этой работе построена система нечеткого поиска, которая может находить слово в списке, даже если оно искажено (содержит опечатки). Для реализации этой системы используется искусственная нейронная сеть Хемминга [1,2].

На вход поисковой системы подается слово *v*, которое будем искать, и текстовый файл со списком слов *V*, в котором будем осуществлять поиск. На выходе получаем - номер *n(w,V)* слова *w* в списке *V*, которое наиболее близко к исходному слову *v*.

```
. . .
'ф'  "00001"   'с'  "01000"   'л'  "10110"
'ы'  "00011"   'а'  "01001"   'щ'  "10111"
. . .
```

Рисунок 1: Коды символов

Входное слово из букв русского алфавита преобразуется в слово в алфавите *01*, которое затем подается на вход нейронной сети, т.е. каждой букве ставится в соответствие слово из символов *0* и *1* длинны *5*. Кодирование строится таким образом, что бы стоящие рядом на компьютерной клавиатуре символы имели близкие по Хеммингу коды (рис.1). Таким образом должно достигаться наиболее эффективное исправление опечаток[3].

тест	*выход*
Акмьлу	Акмола
Алмуты	Алматы
Архунгельск	Архангельск
. . .	. . .
Тьльятти	Тольятти
Тклу	Тула
Черкуссы	Черкассы
Ярьслувль	Ярославль

Рисунок 2: Результат работы программы

В этой реализации размерность распределительного слоя фиксирована и составляет *125* нейронов. Размерность скрытого слоя определяется количеством слов в списке, в котором осуществляется поиск. Размерность выходного слоя равна размерности скрытого слоя нейронной сети.

Реализация сети показала хорошие результаты по обобщению данных, она корректно нашла все, предложенные ей, *112* тестовых слов. Результат работы программы на рис.2.

Необходимо отметить один существенный момент, связанный с архитектурой сети Хемминга. Система хорошо исправляет опечатки, гораздо хуже дело обстоит с пропусками и лишними символами. Хемингово расстояние в этом случае может оказаться слишком большим. Для того, чтобы сгладить этот недостаток, можно подавать на вход как само искомое слово, так и это же слово, исключая по очереди по одному символу в каждой позиции и добавляя по одной букве в каждую позицию. Такой подход позволит найти практически все случаи ошибок - опечатка, пропуск символа, лишний символ[3].

Литература

1. Lipman R. An introdaction to computing with neural nets // IEEE Acoustic,Speech and Signal Processing Magazine,1987,no 2,L.4-22.
2. В.А.Головко, под ред. проф. А.И.Галушкина Нейронные сети: обучение, организация и применение. - Москва : ИПРЖР, 2001
3. А.Арустамов, А.Стариков Ассоциативная память.Применение сетей Хемминга для нечеткого поиска. - http://basegroup.ru/neural/assoc.htm