Document Retrieval and Protein Sequence Matching using a Neural Network

Björn Levin Anders Lansner
blevin@sans.kth.se ala@sans.kth.se
SANS, NADA, Royal Institute of Technology, S-100 44 Stockholm, Sweden

During the development of algorithms for the generation of higher order complex units in neural networks we have come to be interested in free text document retrieval and protein sequence matching. Document retrieval, as it is percieved here, consists of returning a list of the documents in the library that are the most relevant according to a description of a subject, sorted according to descending probability of relevance. Obviously, the key is having a reasonable measure of similarity between the individual documents and a given concept, a measure here provided by the neural network. [1] contains an overview of similar work.

Protein sequence matching is regarded as a special case of document retrieval. What makes it particularly interesting is the existence of generally accepted automatic (but resource consuming) procedures for determining the degree of similarity between sequences or parts of sequences. We have compared our results with those obtained using a system also running on the CM-200 at KTH, based on editing distance without swaps on a database containing 20 000 sequences [2].

In all cases our similarity measures are based on using binary sensors that indicate the existence of short sequences of characters in a document, i.e. feature detectors. These sensors are selected according to quality measures of the type in the table below. The similarity measure is then built up as a weighted sum over the sensors when applied to the search seed. These weights in turn, are computed during a learning phase when the documents themselves are fed to the sensors.

The evaluation of the sensor quality measures has been done on the 8K Connection Machine CM-200 at KTH. These evaluations, done both by comparing rankings of protein sequences with the rankings of the editing distance based system as well as manual examination and testing using plain text documents, showed that especially measure #6 below performed quite well in returning an adequate ranking list of documents or protein sequences.

The ranking is also quite fast. Matching a sequence of 600 characters against a data base of 20 000 documents takes 3.6 s. The matching is then performed using a network that has been trained previously and has been read from the data vault. The training, i.e. selection of sensor units and setting up of the weights of the network, takes approximately 1.5 hours. It should be pointed out that after the training has been completed any number of matchings can be done. It should also be pointed out that a major part of the time is used selecting the sensors, a task that only requires a representative subset of the data. It is also possible to include new documents in the data base very quickly if the material is relatively homogeneous. More details can be found in [3].

#1	$\frac{P_s(\mathtt{ABC}\ldots)}{P_s(\mathtt{A})\cdot P_s(\mathtt{B})\cdot P_s(\mathtt{C})\cdots}$	#4	$\sqrt[n]{P_s(\mathtt{ABC}\ldots)} - \sqrt[n]{P_s(\mathtt{A})\cdot P_s(\mathtt{B})\cdot P_s(\mathtt{C})\cdots}$
#2	$P_s(\mathtt{ABC}\ldots) - P_s(\mathtt{A})\cdot P_s(\mathtt{B})\cdot P_s(\mathtt{C})\cdots$	#5	$\frac{P_s(\mathtt{ABC}\ldots)}{\max\left(\max_X P_s(X\mathtt{ABC}\ldots), \max_X P_s(\mathtt{ABC}\ldots X)\right)}$
#3	$\frac{(P_s(\mathtt{ABC}\ldots))^n}{P_s(\mathtt{A})\cdot P_s(\mathtt{B})\cdot P_s(\mathtt{C})\cdots}$	#6	$\frac{\sqrt{2}\cdot P_s(\mathtt{ABC}\ldots)}{\sqrt{\sum_X P_s(X\mathtt{ABC}\ldots)^2 + \sum_X P_s(\mathtt{ABC}\ldots X)^2}}$

Some of the quality measures tried for selecting sensors. $P_s(\mathtt{ABC})$ stands for the probability that the sensor will be active at a random position while n is its size.

[1] Doszkocs T., Reggia J., and Lin X.: Connectionist Models and Information Retrieval; *Annual Review of Information Science and Technology (ARIST)*, Vol. 25 pp 209 - 260 (1990).
[2] Wallin E.: Optimized Sequence Matching on the CM-2; Masters Thesis, Royal Inst. of Technology, Sweden (1992).
[3] Levin, B. & Lansner, A.: Document Retrieval, Protein Sequence Matching and Sensor Selection Methods using a Neural Network; Royal Inst. of Technology, Tech. rep. TRITA-NA-P9238 (1992)

S. Gielen et al. (eds.), *ICANN '93*