A Neural Network for Probabilistic Information Retrieval

K. L. Kwok

Dept. of Mathematics and Computer Science
Western Connecticut State University
Danbury, CT 06810

ABSTRACT

This paper demonstrates how a neural network may be constructed, together with learning algorithms and modes of operation, that will provide retrieval effectiveness similar to that of the probabilistic indexing and retrieval model based on single terms as document components.

1. Introduction

Since the early 1980's the neural network (NN) or connectionist model as a method of information processing has experienced an explosive revival of interest by researchers of various disciplines [see for example Hinton & Anderson 81, Feldman & Ballard 82, Hopfield 82, Rumelhart & McClelland 86 and references thereof]. This approach stems from an effort to mimic the structures and operations of the brain, and it is hoped that some of the human-like behavior such as vision, language understanding, hearing, etc. that are difficult to solve by conventional artificial intelligence (AI) methods may find plausible explanation with this new paradigm. Information retrieval (IR) deals with large collections of textual material, and its aim is to satisfy user queries and needs (also expressed in natural language text) with documents that are relevant. This conceptual matching belongs to the category of difficult problems and it is therefore interesting to investigate whether this NN approach may be of use to IR. During the past twenty or so years, a large body of research effort has gone into IR, and a certain level of theoretical understanding and experimental performance has been achieved with traditional retrieval methods such as the vector and generalized vector models [Salton 68, Salton et.al. 83, Wong et.al. 86] and the recent latent semantic indexing work [Deerwester et.al. 88], Boolean and Fuzzy Set models [Radecki 79, Waller & Kraft 79, Bookstein 81], and the probabilistic models [Bookstein & Swanson 75, Yu & Salton 76, Robertson & Sparck Jones 76, van Rijsbergen 79, Croft 83, Kwok 88]. It therefore appears reasonable to expect that an NN approach should at least provide such a level of performance as a base. This paper represents an attempt to employ the NN paradigm to reformulate the probabilistic model of IR with single term as document components [Kwok 85, 87, 88], since it has a sound theoretical foundation and has also been experimentally determined to be comparable or better in effectiveness than some of the best methods available.

From this basis we hope that further developments may lead to better results. Other people have applied the NN approach to IR [Mozer 83, Belew 86, Brachman & McGuinness 88] previously, but they have in general not taken account of what has been accomplished by information retrieval researchers in recent years. In [Salton & Buckley 88] similar effort has also been attempted.

2. Neural Networks

Neuro-scientists have studied the brain for many years and have found that it is composed of a vast number (billions) of cells of different types called neurons [Llinas 75]. Each neuron connects (fan-out) to, as well as is connected (fan-in) from, many other neuron cells via synaptic junctions. Typically these connections may be in the thousands. Some neurons are connected to our sensory organs and can be activated by external environmental stimuli. This activity in a cell, if it exceeds a certain threshold, can generate an output (i.e. the cell fires) affecting in parallel other neurons connected to it. These other neurons may also fire, and lead to a spreading of activation, representing a processing of the information. The state of activation of certain subsets of the neurons is then taken as the brain's response to the stimuli, and may lead to certain motor actions. This processing is parallel and distributed and the model is quite different from the sequential, instruction by instruction processing of a computer. Another important aspect of the brain is that it can learn from experience, and this learning is assumed to be reflected as changes in the connection strengths between neurons, namely, the synaptic connections are plastic [Hebb 49, Marr 69]. The above is a very brief summary of the structure and operations of the brain. What follows is a description of the usual hypothetical model of neurons, neural networks and their functions that cognitive psychologists, engineers and computer scientists use and that are of interest to us for applications in IR.

2.1 Single Neuron

A single neuron j is regarded as a hypothetical processing unit as shown in Fig. 1. It receives input from (i = 1 .. n) other neurons via synaptic connections of strengths w_{ji}. In addition, it may receive input from the environment represented as E_j, and it may also have a constant bias B_j of its own. The resultant net input for neuron j is then usually taken as a linear sum of all the possible inputs, thus:

$$net_j = \Sigma_i \, w_{ji} * o_i + E_j + B_j \qquad (1)$$

where o_i is the output from neuron i, affecting j. This net_j input may lead to a new activity level on neuron j governed by an activity function F:

$$a_j = F(a_j, net_j) \qquad (2)$$

This new activity on j, if sufficiently intense, may in turn

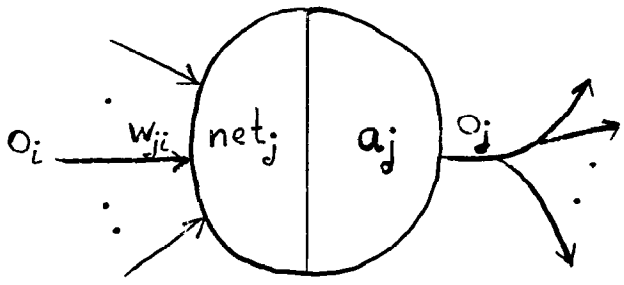


Fig.1: Inputs, activity and outputs of neuron j.

trigger its firing and carry signal to those connected from j. The output of j may be governed by an output function f:

$$o_j = f(a_j) \qquad (3)$$

Typical examples of the output functions f are the identity function or a (nonlinear) threshold. Examples of the activity function F are the sigmoid function or various linear functions in net_j and a_j, including the identity [Rumelhart & McClelland 86 (Ch.2)].

There can be many assumptions for neuron activities. For example, activation values can be continuous, continuous with upper and lower bounds, graded, or just binary (1 or 0). The activation levels may also be interpreted as a probability of the cell firing ($o_j = 1$) or not ($o_j = 0$), leading to a stochastic model. Connection strengths are real-valued, but may be positive (excitatory) or negative (inhibitory).

2.2 A Network of Neurons

A set of previously described neurons interconnected together form a neural network of arbitrary complexity. The usual architecture is to organize the neurons into one or more layers. Many dynamics of activation spreading can be considered. For example, activation spreading may be from an input layer towards the output layer (feed-forward) or vice versa (feed-backward), or in both ways (interactive models), or the net may be allowed to relax to an equilibrium state. Neuron activities may also be updated synchronously or asynchronously.

How the connection strengths are assigned is the major concern of a neural network. Connections may be given initial values suitable for explaining a problem such as the model for the Necker Cube perception [Feldman & Ballard 82], but the basic idea is that these connections are adaptive and that the network should be able to learn from experience or feedback to accomplish or explain a certain goal. An example is the back-propagation algorithm [Rumelhart, Hinton & Williams 86] which varies the connection strengths by seeking to minimise the mean square error between teaching signals and the outputs of a net. The specification of a learning algorithm for these connections is therefore a crucial characteristic of a particular NN model.

For our purposes we will use a simple 3-layer architecture, as shown in Fig. 2. Neurons in layer Q (later to be identified with a set of queries) may receive external input and are connected to the neurons in layer T, the hidden units (index term nodes). Layer T is then connected to layer D (document nodes), which may serve as output units. As an initial attempt we will keep the net simple and disallow connections within layers. The connections between layers are directed but of unequal weights. We will also consider activation to spread both forward and backwards separately. The operation of the net, as well as its special learning algorithms, will be explained in Section 4 after we review the probabilistic indexing and retrieval model.

3. Probabilistic indexing and Retrieval using Single Terms As Document Components

Probabilistic retrieval aims at providing an optimal ranking of a document collection with respect to a query. This is based on a decision function, which for ranking purposes may be transformed via the Bayes' Theorem

into the following more useful form: $P(d_i|+R)/P(d_i|-R)$, which is the ratio of the conditional probabilities that given relevance (+R) or non-relevance (-R) to a query that one finds documents of the description d_i. The theory as first proposed in [Robertson & Sparck Jones 76, van Rijsbergen 79] makes two basic assumptions: (a) the index terms used for characterizing a document or query are independent, and (b) the document index term descriptions are binary. The query index term descriptions however are assigned weights, and they are the above ratio for each term when they are regarded as independent. To estimate these probabilities one needs a sample of relevant documents, for example those relevant ones obtained from a relevance feedback operation. (The sample of non-relevant documents are usually taken as the whole collection minus the relevant sample, and is a very good estimate because there are usually few relevant documents per query.) Once the samples are identified, occurrence statistics of each index term in the samples can be counted and the probabilities estimated. The resulting weight assigned to each term k of query a is then given by:

$$g_{ak} = g^r_{ak} + g^s_{ak} = \log [r_{ak}/(1-r_{ak})] + \log [(1-s_{ak})/s_{ak}] \quad (4)$$

where r_{ak} = P(term k present|+R) and s_{ak} = P(term k present|-R). They are the conditional probabilities that given relevance (+R) or non-relevance (-R) to query q_a that term k will be found present in the documents. If document d_i has the vector $(x_{i1},.. x_{ik},..)$ to denote the presence (x_{ik}=1) or absence (x_{ik}=0) of term k as its representation, then the optimal weight to be attached to d_i for ranking purposes is:

$$W_i = \sum_k x_{ik}\, g_{ak} \quad (5)$$

which sums over all terms common to d_i and q_a.

The probabilistic retrieval model is theoretically sound and also provides good retrieval results. However, it is not without its drawbacks. For example, it assumes that document index term descriptions are binary, and therefore does not make full use of the information available in the within-document term frequencies. It has a term weighting problem at the initial stage since one has to obtain a sample of relevant documents for each query first. It also ignores the reverse situation of probabilistic indexing [Maron & Kuhn 66], where one can use a document as a focus and ask which queries are relevant to it. These and some other considerations can be alleviated to a certain extent by the approach proposed in [Kwok 85,87,88], where each document is viewed as constituted of many conceptual components, and where one works in a universe of document components instead of a collection of documents. The components are assumed to be independent and unambiguous in meaning. They may be phrases, for example, but for this investigation we take them as single terms. The result of this extension is that all within-document term frequencies are made use of effectively. The theory can self-bootstrap because every document has a relevant component sample set to start with, namely, its own self. In addition, by consideration of the situation of either the query or the document (or both) as the focus, probabilistic weighting of query terms and document terms can be accommodated, leading to asymmetric as well as symmetric weighting formulae of the type proposed in [Robertson et.al. 86]. The resultant formulae may be summarized as follows. When one treats a query q_a as the focus and

asks whether a document d_i is relevant or not, we have the weight assigned to each query term k as in (4), but with:

$$r_{ak} = q_{ak}/L_a, \quad s_{ak} = F_k/N_w \qquad (6)$$

where r_{ak} in our case of counting terms as components can be expressed as the k-th term frequency q_{ak} divided by the length L_a of the query. s_{ak}, which is the probability of occurrence of term k given non-relevance, can be estimated by the collection frequency F_k of term k divided by the size of the component universe N_w, which is a count of all the terms used. In this case of treating the query as the focus, the weight assigned to document d_i for ranking purposes then becomes:

$$WQ_i = \Sigma_k \, (d_{ik}/L_k) \, g_{ak} \qquad (7)$$

summing over all terms k common to both d_i and q_a. (The Q of WQ reminds us that this weight is obtained with the query as the focus.) On the other hand, when one treats a document d_i as the focus and asks whether the query q_a is relevant or not, we have the weight assigned to each document term k as:

$$g_{ik} = g^r_{ik} + g^s_{ik} = \log [r_{ik}/(1-r_{ik})] + \log [(1-s_{ik})/s_{ik}] \qquad (8)$$

with

$$r_{ik} = d_{ik}/L_i$$
$$s_{ik} = (F_k-d_{ik})/(N_w-L_i),$$

where r_{ik} and s_{ik} have the same interpretation as for r_{ak} and s_{ak}, but for the document d_i instead. d_{ik} and L_i are respectively the term frequency of k and the length of d_i respectively. The weight assigned to document d_i for ranking purposes then becomes:

$$WD_i = \Sigma_k \, (q_{ak}/L_a) \, g_{ik} \qquad (9)$$

again summing over all terms k common to both d_i and q_a. Finally, we may also introduce the symmetric formula which is the sum of the two:

$$W_i = \Sigma_k \, [(q_{ak}/L_a) \, g_{ik} + (d_{ik}/L_i) \, g_{ak}] \qquad (10)$$

This kind of <u>initial</u> weighting has also been called indexing based on a principle of document self-recovery [Kwok 86,88], because it provides an optimal weighting of the terms if we regard an item (document or query) as a 'query' and require that this 'query' should retrieve itself optimally in the universe of components. In the next sections, we will show how a neural network may be constructed with operations that can provide similar performance.

4. A Neural Network Approach to Information Retrieval

As discussed earlier, our current goal is to design a NN that will default to the IR retrieval results of Section 3. We will divide this section into four parts as follows: 1) the architecture of the net; 2) initial connection strengths; 3) mode of operation for retrieval; and 4) learning algorithms and initial learning. We are interested in NN as a new method of general information processing and its application to IR, rather than claiming this to be the way how the brain does document retrieval.

4.1 Architecture of the Net

The 3-layer (Q, T, D) network discussed in Section 2.2 and shown in Fig. 2 is interpreted as queries connected to index terms to documents. These connections are bi-directional and asymmetric. Queries and documents are regarded as neurons of the same category, and can play either as input or output units. Intra-layer connections are disallowed

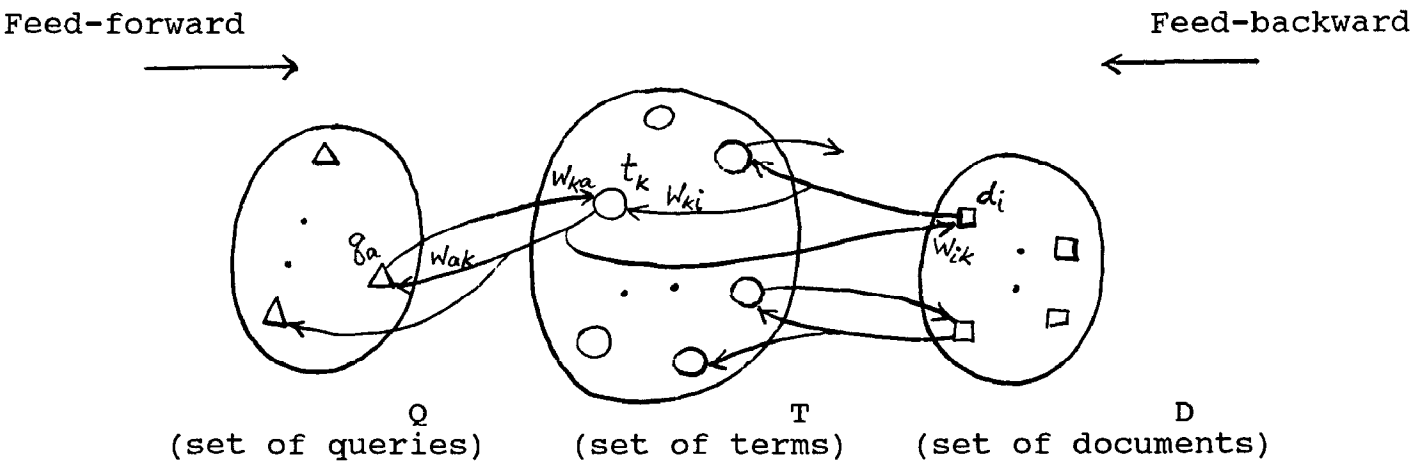


Fig.2: 3-Layer Net for IR (not all connections are shown)

in this report. We will also assume that the activities on the neurons take on continuous values, and both the output and activation functions are taken as the identity function.

4.2 Initial Connection Strengths

The connection strengths between neurons may be initially assigned in the following manner. (In Section 4.4, we will propose how they may be acquired from scratch.) From a query neuron a (document neuron i) to an index term neuron k, the connection strength will be $w_{ka} = q_{ak}/L_a$ ($w_{ki} = d_{ik}/L_i$). The interpretation is that these strengths represent the inference that given the presence of this query a (or document i), it will have a probability q_{ak}/L_a (d_{ik}/L_i) of using the constituent term k, and is obtained from the manifestation of the query (document) text. From a term neuron k to a query neuron a (or document neuron i), the connection strength w_{ak} (w_{ik}) will be taken as composed of two parts: $w^r_{ak} + w^s_{ak}$, ($w^r_{ik} + w^s_{ik}$) and these are to be identified with $g^r_{ak} + g^s_{ak}$ of Eqns. (4,8), Section 3. w^s_{ak} as discussed before, is the estimate from a sample of non-relevant documents and can be quite accurately estimated by the following: $w^s_{ak} = w^s_{ik} = \log (1-s_k)/s_k$, where $s_k = F_k/N_w$. This represents the log-odds evidence that if term k is used, it will be dealing with the contents of query a or document i to this extent. It is a property of term k only, and gets modified as the term usage in the net changes. w^r_{ak} and w^r_{ik} are assigned initial values $\log[p/(1-p)]$, with p being a small positive constant, the same for all terms. Together, w_{ak} (w_{ik}) provides a discrimination based, inverse-document-frequency (IDF) type of weighting for the case of single terms as document components, and represents the best known information of the usefulness of term k without any further content-oriented information. How retrieval may be done is presented in the following sub-section.

4.3 Modes of Operation for Retrieval

Once the net has been initialized as in Section 4.2, we may view retrieval as the result of a feed-forward or feed-backward spreading of activation. For example, analogous to the query-weighted formula WQ_i of Eqn. (7) with query q_a in attention as the focus, we will assume that each document d_i will have its activity clamped to

a value of 1 in turn. This activity then spreads to the term neurons and to q_a via connections of non-zero strengths. Each document may then be evaluated for relevance to q_a (or not) based on whether the activity received at q_a exceeds a certain predetermined threshold value. For comparison with traditional evaluation methods, we will however use this activity for ranking the documents. In the reverse situation, we can clamp the activity of query q_a to 1, spread its activity towards each document in attention (d_i), and use the activity received by d_i for ranking its relevancy to q_a. This mode corresponds to the document-weighted formulation WD_i of Eqn. (9). If we add the corresponding activities of the above two cases, and use the sum for relevancy ranking, we recover the symmetric sum formulation of Eqn. (10).

As an example of such a retrieval, we may consider a query neuron a being clamped to an activity of 1. Any term neuron k affected by the query will receive input $net_k = q_{ak}/L_a$ (i.e. $1 * w_{ka}$), leading to an activity and output signal of the same value. During the next time step, a document neuron d_i in attention and having terms k in common with the query will receive as input:

$$\sum_k (q_{ak}/L_a)\, w_{ik} = \sum_k (q_{ak}/L_a) * \log [p/(1-p) * (N_w-F_k)/F_k].$$

This again will be reflected as an activity on di, and experiments have shown [Kwok xx] that they provide much better results than the traditional, document-mode IDF weighting first proposed by (Sparck Jones 72) and popular among experimental retrieval work.

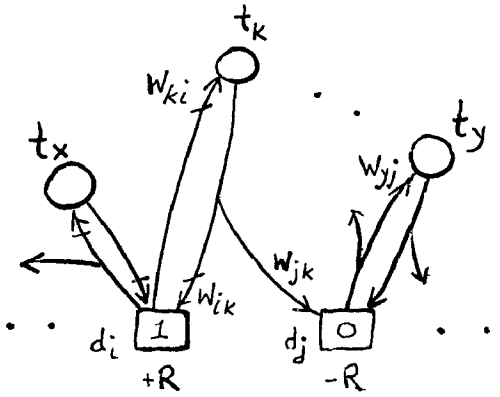
4.4 Learning Algorithms

To improve on the NN, we assume that learning algorithms exist which can lead to changes in the connection strengths. Consider the net to be initially blank. As each document neuron d_i is created and fed into the network one by one, associated new term neurons are also created and d_i would acquire new links between itself and the existing or new term neurons. w_{ki} will be set to d_{ik}/L_i as discussed in Section 3.3, and $w_{ik} = w^r_{ik} + w^s_{ik} = \log [p/(1-p)] + \log [(N_w-F_k)/F_k]$. N_w is some large constant and F_k may be estimated as a function of D_k (i.e. $F_k(D_k)$), which is the current fan-in factor (or document frequency) of neuron k. Thus each time a new link is created at term k (because of a new document), all connections w^s_{ik} emanating from k have to be adjusted as well.

Once the above is established the net can further learn from each document itself, (even though there is no feedback information yet), since we know that d_i must be self-relevant. Consider the document neuron i to be under attention, with activity clamped to 1. It sends signals to the connected term neurons, from which the activity further spreads. Since d_i is also the only known relevant item at this point (see Fig. 3), only the connections w_{ki} and w_{ik} (for all terms k of document i) will be modified according to the following rules:

$$\text{A: } \Delta w_{ik} = h(w_{ik}, a_k) \text{ if neuron i is relevant;}$$
$$= 0 \text{ otherwise.} \quad (11)$$

This may be seen as a type of Hebbian correlational learning [Hebb 49] and is of the same nature as those used in [Belew 86]. It is not truly Hebbian because we would use only the activity on neuron k and not that on i. In contrast to the delta rule [Widrow & Hoff 60, Sutton & Barto 81, Rumelhart & McClelland



d_i clamped to activity 1 & self-relevant, hence (w_{ki}, w_{ik}) are modified, but not w_{jk}, or w_{yj}.

Fig.3: Self-Learning by Document

86], we use relevance information not as a teaching signal to be reproduced, but as confirmation of which connections to modify. What we want to learn, according to the probabilistic model, is an estimate of the occurrence probability of term k in the current relevant set of d_i, and this happens to be residing as an activity a_k on neuron k. Hence the connection strengths would be assumed to learn according to the following:

$$\Delta w_{ik} = \Delta r_{ik}/[r_{ik}*(1-r_{ik})] \quad (12)$$

here r_{ik} is the probability before learning, viz:

$$\exp(w_{ik}-w^s_{ik})/[1+\exp(w_{ik}-w^s_{ik})].$$

We assume that learning takes place gradually over many iterations with a learning rate eta, and the rule for learning of the probability r_{ik} for the v+1 iteration is:

$$r_{ik}^{v+1} = (1-\text{eta})*r_{ik}^{v} + \text{eta}*a_k, \quad 0 < \text{eta} < 1 \quad (13)$$

$$r_{ik}^{0} = \text{initial value of } r_{ik} = p.$$

As v approaches infinity, we see that the learnt probability r_{ik}^{v} approaches a_k. For each iteration step, the change in r_{ik} is therefore:

$$\Delta r_{ik} = \text{eta} * (a_k - r_{ik}) \quad (14)$$

This together with Eqns. (11,12) defines the update function h. For the case of the documents learning its own relevance, we have used eta = 0.5 for all documents, and iterate some 20 times. This is equivalent to setting the initial connection strengths w_{ik} to log $[d_{ik}/(L_i-d_{ik}) * (N_w-F_k)/F_k]$, which is approximately that of the term weight obtained from the principle of document self-recovery (Eqn. 8) [Kwok 87, 88]. This learning process is also assumed to take place for the connections w_{ak} between the terms and query neurons. Thus queries and documents are regarded as items of the same category, we only use different symbols to denote them in the drawings for clarity. When one plays the part of external input, the other assumes the role of output and vice versa. However, queries are usually much more transient, and we have calculated the universe statistics without accounting for the usage frequencies from the queries.

When the link w_{ik} was adjusted, simultaneously the other link w_{ki} may also be assumed to learn, as in Eqns. (13, 14), thus:

$$\text{B:} \quad \Delta w_{ki} = \Delta r_{ki} = \text{eta} * (a_k - r_{ki}) \quad (15)$$

In this case of learning from a document itself however, nothing is changed because the initial values of w_{ki} are exactly those of a_k, the activities on neuron k. This however will not be true when we consider the case of relevance feedback from several documents.

If the net is now used for retrieval, it will perform with

similar effectiveness to those reported in [Kwok 88] using the principle of document self-recovery for initial indexing, and the asymmetric and symmetric formulae of Eqns. (7,9,10).

5. Conclusion and Discussion

We have shown how the probabilistic retrieval formulation using single terms as document components may be implemented in a neural network, together with the necessary learning algorithms. Interpretations of the connection strengths are also given. It achieves optimal ranking based on the probability of relevance for a document with respect to a given query and assumming term independence. An advantage of this net is that it defaults to known results and is amenable to comparisons with previous investigations and evaluations by information retrieval researchers.

This is of course only the first step. More interesting would be the situations when we switch on the interactions among the neurons within a layer. In that case, more sophisticated dynamics of activation such as those employed for associative memory or pattern completion under constraint satisfaction requirements (Hopfield 1982, 1984, Hinton, Sejnowsky & Ackley 1984, Smolensky & Riley 1984) may probably be more appropriate.

6. Acknowledgment

Mr. Alan S.K. Kwong did the preliminary programming work. Trial results have been performed on the Cornell National Supercomputer Facility, a resource of the Center for Theory and Simulation in Science and Engineering at Cornell University, which is funded in part by the National Science Foundation, New York State, and the IBM Corporation and members of the Corporate Research Institute.

References

Belew, R.K. (1986). Adaptive information retrieval: machine learning in associative networks. Ph.D. Thesis, University of Michigan.

Bookstein, A. (1981). A comparison of two weighting schemes for Boolean retrieval. In: Oddy,R.N; Robertson, S.E.; van Rijsbergen, C.J.; Williams, P.W. (ed.): IR Research. London: Butterworths.

Bookstein, A. ; Swanson, D. R. (1975). A decision theoretic foundation for indexing. J. of ASIS. 26:45-50.

Brachman, R.J. & McGuinness, D.L. (1988) Knowledge representation, Connectionism, and Conceptual retrieval. In: Chiaaramella, Y. (ed.): Proc. of 11th ACM Intl. Conf on R&D in IR. Grenoble:PUG.

Croft, W. B. (1983) Experiments with representation in a document retrieval system. Info. Tech.: R&D. 2:1-21.

Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K. (1988) Indexing by latent semantic analysis. In: Chiaarmella, Y. (ed.): Proc. of 11th ACM Intl. Conf. on R&D in IR. Grenoble: PUG.

Feldman, J.A. & Ballard, D.H. (1982). Connectionist models and their properties. Cognitive Science, 6, 205-254.

Hebb, D.O. (1949). The organization of behavior. N.Y.:Wiley.

Hinton, G.E. & Anderson, J.A. (Eds.) (1981). Parallel models of associative memory. Hillsdale NJ: Erlbaum.

Hinton, G.E.; Sejnowsky, T.J.; Ackley, D.H. (1984). Boltzmann Machines: constraint satisfaction networks that learn. T.R. CMU-CS-84-119. Pittsburgh: Carnegie-Mellon Univsity.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective comput-

ational abilities. Proc. Natl. Acad. Sci. USA, 79, 2554-2558.
Hopfield, J.J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Natl. Acad. Sci. USA, 81, 3088-3092.
Kwok, K.L. & Kuan, W. (1988). Experiments with document components for indexing and retrieval Info. Mngmt. Proc., 24, 405-417.
Kwok, K. L. (1986). An interpretation of index term weighting schemes based on document components. In: Rabitti, F. (ed.): Proc. of 1986 ACM Conf. on R&D in IR. Baltimore: ACM, 275-283.
Kwok, K.L. (1985). A probabilistic theory of indexing and similarity meausre based on cited and citing documents. J. of ASIS. 36:342-351.
Kwok, K.L.(19xx). Further results of probabilistic retrieval based on single terms as document components. (to be submitted.)
Llinas, R.R. (1975). The cortex of the cerebellum. Scientific American, 232, 56-71.
Marr, D. (1969). A theory of cerebellar cortex. J. of Physiology, 202, 437-470.
Mozer, M.C. (1984). Inductive information retrieval using parallel distributed computation ICS T.R. 8406. La Jolla: UCSD.
Radecki, T. (1979). Fuzzy-set theoretical approach to document retrieval. Info. Proc. Mgmnt. 15:247-259.
Robertson, S. E.; Maron, M.E.; Cooper, W.S. (1982). Probability of relevance: a unification of two competing models for document retrieval." Info. Tech.:R&D 1:1-21.
Robertson, S.E.; Sparck Jones, K (1976). Relevance weighting of search terms." J. of ASIS. 27: 129-146.
Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. (1986). Learning representations by back-propagating errors. Nature 323, 533-36.
Rumelhart, D.E. & McClelland,J.L. (1986). Parallel distributed processing, Vol. I Foundations & Vol. II Psychological and biological models. Cambridge, MA: MIT Press.
Salton, G. & Buckley, C. (1988). On the Use of Spreading Activation Methods in Automatic Information Retrieval. T.R. 88-907, Ithaca: Computer Science Dept., Cornell University.
Salton, G; Fox, E.A; Wu, H(1983). Extended Boolean information retrieval. Comm. of ACM. 26: 1022-1036.
Salton, G. (1968). Automatic information organization and retrieval. New York: McGraw Hill
Smolensky, P. & Riley, M.S (1984) Harmony theory: problem solving, parallel cognitive models, and thermal physics. ICS T.R. 8404. La Jolla: UCSD.
Sparck Jones, K. (1972). A statistical interprestation of term specificity and its application in retrieval. J of Doc. 8:11-21.
Sutton, R.S.& Barto, A.G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. Psych. Rev 88:135-170.
van Rijsbergen, C.J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. J. of Doc. 33: 106-119.
Waller, W.G; Kraft, D.H (1979). A mathematical model for a weighted boolean retrieval system. Info. Proc. Mgmnt. 15: 235-245.
Widrow, G.& Hoff, M.E. (1960). Adaptive switching circuits. IRE WESCON Convention Record, Part 4 pp.96-104.
Wong, S.K.M; Ziarko, W; Raghavan, V.V.; Wong, P.C.N. (1987). On modeling of information retrieval concepts in vector spaces. TODS, 12, 299-321.
Yu, C. T.; Salton, G. (1976). Precision weighting - an effective automatic indexing method." J. of ACM. 23:76-86.