Associative Database for Information Retrieval

Sándor **Dominich**

Department of Computer Science, University of Veszprem, Egyetem u. 10, 8200 Veszprem, Hungary, E-mail: dominich@dcs.vein.hu

An associative database model (*AI*^{2}*R*) *is defined, using Artificial Neural Networks* (*ANN*), *based on Interaction Information Retrieval* (*I*^{2}*R*). *Retrieval is preceded by an interaction between query and database, and means recalled local memories. It is shown that AI*^{2}*R's underlying abstract mathematical structure is a matroid and that AI*^{2}*R is in the* **NP**-*Class. An implementable model is worked out based on state equations, and an application is briefly described.*[1]

1. Associative Interaction *IR* (*AI*^{2}*R*)

1.1. The *AI*^{2}*R* Model

The Interaction Information Retrieval (*AI*^{2}*R*) model was firstly introduced in [2], [3]. Given a set $D = \{d_1, d_2,..., d_i,..., d_M\}$ of *documents*, $M > 1$. An Artificial Neural Network (*ANN*) *D-Net* $= \langle \aleph, W, L \rangle$ is associated as follows: $\aleph = \{\nu_i$: ν_i artificial neuron assigned to d_i, $i = 1, 2, ..., M\}$ denotes a set of neurons, L: $\aleph \times \aleph \rightarrow \mathbf{R}_+^{K_{ij}}$, $L(\nu_i, \nu_j) = \mathbf{w}_{ij} = (w_{ij}^1,..., w_{ij}^k,..., w_{ij}^{K_{ij}}) \in \mathbf{R}_+^{K_{ij}}$ denotes connection strengths or weights, $i, j = 1, 2,..., M$, and $W = \{\mathbf{w}_{ij}$: $i, j = 1, 2,..., M, i \neq j\}$ denotes a set of weights. The following conditions hold: $\mathbf{w}_{ij} \neq \mathbf{0} \Rightarrow \mathbf{w}_{ji} \neq \mathbf{0}$, $\forall i, j$; $0 \leq w_{ij}^k \leq 1$, $\forall i, j, k$. The *state* of ν_i is denoted by z_i. An *activation spreading* takes place in *D-Net*.

Definition 1. *Feeding* a new artificial neuron ν into a *D-Net* $= \langle \aleph, W, L \rangle$ means obtaining a new *D'-Net* $= \langle \aleph', W', L' \rangle$ where $\aleph' = \aleph \cup \{\nu\}$, L': $\aleph' \times \aleph' \rightarrow \mathbf{R}_+^{K_{ij}}$.

W' contains more weights than W hence $|W'| > |W|$, $W' \setminus W \neq \varnothing$. One can distinguish three cases: a) W' contains the new weights for the fed ν and all of the old weights unchanged, b) W' contains the new weights for the fed ν and all of the old weights changed, c) W' contains the new weights for the fed ν and both changed and unchanged old weights.

[1] Research partly supported by Research Grant Pro Renovanda Cultura Hungariae 98/31.

Definition 2. Let $\langle \aleph, W, L \rangle$ be a *D-Net* and $\langle \aleph', W', L' \rangle$ be the fed *D'-Net.* The set difference $W' \setminus W$ is *interaction* I between ν and *D-Net*, $I = W' \setminus W$. When only a) occurs, interaction is called a *pseudo-interaction*: $I = W' \setminus W = \{\mathbf{w}'_{\nu j}, \mathbf{w}'_{j\nu}, \forall j\}$. When b) or c) or both occur, interaction is called a *real interaction*: $I = W' \setminus W = \{\mathbf{w}'_{\nu j}, \mathbf{w}'_{j\nu}, \exists i, j: \mathbf{w}'_{ij} \neq \mathbf{w}_{ij}\}$.

Definition 3. Let $\langle \aleph, W, L \rangle$ be a *D-Net* and $\langle \aleph' = \aleph \cup \{\nu_q\}, W', L' \rangle$ the associated fed *D'-Net*. *Interaction Information Retrieval* (*IIR=I²R*) on *D'* is a 2-tuple $\langle D', \Re \rangle$ where $\Re(q) = \{d_i: \nu_i$ winner in activation spreading started at $\nu_q\}$ is the set of retrieved documents in response to *q*.

Definition 4. A *reverberative circle* ς is a sequence $\varsigma = \nu', ..., \nu^p, ..., \nu^V$ of artificial neurons where: $\nu' = \nu^V$ and ν^p is a winner, i.e. is the most active of all elements succeeding its predecessor, i.e. ν^p such that $z_p = \max_j \{z_j: \mathbf{w}_{p-1,j} \neq \mathbf{0}, \ p - 1 \neq j\}$.

Definition 5. An element ν *recalls* a reverberative circle ς if ς is formed due to an activation spreading originating at ν.

Finiteness in practical implementation is guaranteed by the following:

THEOREM 1. *There exists at least one reverberative circle ς in a D-Net recalled by a non-isolated element* ν. ∈

Definition 6. Let $\langle \aleph, W, L \rangle$ be a *D-Net* and $\langle \aleph', W', L' \rangle$ the associated fed *D'-Net*. *Associative Interaction Information Retrieval* (*AI²R*) on *D'* is a 2-tuple $\langle D, \Re \rangle$ where $\Re(q) = \{d_i : \nu_i \in \varsigma, \ \varsigma$ recalled by non-isolated $\nu_q\}$.

Specific properties of *AI²R* are investigated in [1], [4]–[14].

1.2 Mathematical properties of *AI²R*

As a new result, it is now shown that:

THEOREM 2. *Retrieval in AI²R means defining a matroid.*
Proof. $\langle \aleph', W', L' \rangle$ can be assigned a complete, directed, weighted multigraph *G* as follows: i) each artificial neuron ν_i is assigned a vertex v_i, ii) there are two oppositely directed edges (opposite arcs), e_{ij} and e_{ji}, between every pair of vertices v_i and v_j ($i \neq j$) having weights u_{ij} and u_{ji} respectively, where $u_{ij} = \sum_{k=1}^{K_{ij}} w_{ij}^k$ and $u_{ji} = \sum_{k=1}^{K_{ji}} w_{ji}^k$. Any reverberative circle ς corresponds to a circle *C* in graph *G*. Let $N = \{\nu_\alpha: \alpha = 1, 2,..., A\}$ denote the artificial neurons

traversed before ς is recalled by ν_q. Then N corresponds to a path $P = \{\nu_\alpha: \alpha = 1, 2,..., A\}$. This means that retrieval defines a connected subgraph H with circles and cutpoints. Hence a block-cutpoint graph T can be assigned to subgraph H which generates a matroid.

Thus, the ideal retrieval case is defined as follows:

Definition 7. An AI^2R is *optimal* if $\Re(q)$ is the maximal matroid.

As another new result, it is now shown that the computational complexity of AI^2R is as follows:

THEOREM 3. *AI^2R is in the* **NP**-*Class*.
Proof. The number $s(M, p)$ of evaluations in the search space is $s(M, p) = M\cdot(M - 1 + \sum_{p=1}^{M} C_M^p)$ where M denotes the number of elements and p is the maximum number of most active elements. Because $s(M, p) = M\cdot(M - 1 + \sum_{p=1}^{M} C_M^p) = M\cdot(M - 1 + 2^M - 1) = O(2^M)$, AI^2R is in the **NP**-Class.

2. Implementable Model of AI^2R

Activity level z_i can be expressed using the General Network Equation:

(1) $$\frac{dz_i(t)}{dt} = I_i(t) - z_i + \Sigma_j T_{ij} f_j(z_j(t))$$

where z_i denotes the *activity level* of element ν_i , $I_i(t)$ denotes the external input to ν_i, t denotes time, $f_j(z_j(t))$ denotes the influence of ν_j's activity level on z_i, and T_{ij} denotes connection strengths between ν_i and ν_j. Because there is no external input we take $I_i(t) = 0$. T_{ij} corresponds to $\mathbf{w}_{ij}$, and it can be taken as e.g. the sum of all weights. The term $f_j(z_j(t))$ is taken as unity as it is assumed that all elements have equal influence. Thus the network equation becomes:

(2) $$\frac{dz_i(t)}{dt} = - z_i + \Sigma_j \Sigma_k w^k_{ji}$$

Let ν_q denote an artificial neuron under focus. An activation is started and spread from ν_q at time t_0 by clamping its state to 1, i.e. $z_q(t_0) = 1$, and $z_i(t_0) = 0$ $\forall i \neq q$. At $t_1 > t_0$ the maximum activity *max* z_i is to be found from among all i, z_i being influenced by z_q, i.e. $j = q$. z_i is given by the solution of the following Cauchy-problem:

(3) $dz_i(t)$

$$\frac{dz_i}{dt} = -z_i + \Sigma_k w_{ji}^k, \qquad z_i(t_0) = 0, \quad j = q$$

Denoting the term $\Sigma_k w_{ji}^k$ by s_i, the solution is $z_i(t) = s_i\, e^{-t} (e^t - 1)$. Assuming now that ν has the maximum activity z of all its 'competitors' ν_i, we have:

(4) $$z(t) \geq z_i(t) \Leftrightarrow (s - s_i)\, e^{-t} (e^t - 1) \geq 0$$

Because $e^{-t} (e^t - 1)$ is positive it follows that $s \geq s_i$. In other words, maximum activity is equivalent to maximum weights sum $\Sigma_k w_{ji}^k$.

Example. Each document d_i is associated a vector $\mathbf{t}_i = (t_{ik})$, $k = 1, ..., n_i$, of identifiers. There are two pairs of links per direction. The one is the frequency of a term given a document, i.e. the ratio between the number f_{ijp} of occurrences of term t_{jp} in object d_i and the length n_i of d_i, i.e. total number of terms in d_i:

(5) $$w_{ijp} = \frac{f_{ijp}}{n_i}, \quad p = 1, ..., n_j$$

The other is the extent to which a given term reflects the content of a document, i.e. the inverse document frequency. f_{ikj} denotes the number of occurrences of term t_{ik} in d_j, df_{ik} is the number of documents in which t_{ik} occurs, w_{ikj} is given by the inverse document frequency formula, and thus represents the extent to which t_{ik} reflects the content of d_j:

(6) $$w_{ikj} = f_{ikj} \log \frac{2M}{df_{ik}}$$

The other two connections - in the opposite direction - have the same meaning as above: w_{jik} corresponds to w_{ijp}, while w_{jpi} corresponds to w_{ikj}. Input s_i of d_i is defined as the sum:

(7) $$\sum_{p=1}^{n_j} w_{jpi} + \sum_{k=1}^{n_i} w_{jik}$$

The implementable model was first tested in Mathcad 8 Professional Plus. A typical activities surface of interconnected documents (before interaction) is shown in Figure 1. Figure 2 shows the activities surface after the query has interacted with documents: previous documents structure has changed.

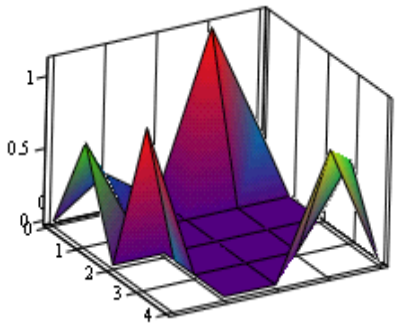

Figure_1

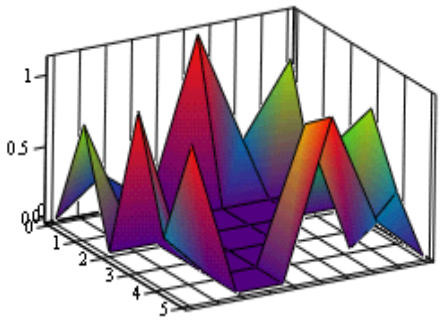

Figure_2

3. *AI²R* Application

A departmental *IR* database system to search literature on concurrency control in Distributed Database Systems (*DDBS*) was build using *AI²R*. The application runs under Windows 95 in current style (self setup, windows operated, etc.) and was written in C++. It has a modular structure as follows: Object Editor (*OE*), Database Editor (*DBE*) and Retrieval (*R*).

OE (Figure 3) allows for creating/editing objects, i.e. documents + corresponding identifiers.

Figure 3. Screen of Object Editor (*OE*). The upper scrolling window allows for creating (e.g. by typing, pasting, opening) a document content (text + images) and saving. The lower scrolling window allows for creating, modifying and deleting identifiers of document content. There also are online help and exit functions. *OE* is mouse operated and used offline.

DBE (Figure 4) allows for collecting objects into databases as well as updating databases.

Figure 4. Screen of Database Editor (*DBE*). The upper scrolling window allows for flexibly collecting/deleting objects into/from databases. An existing database can be open for update. Objects names in a database appear here. The lower scrolling window shows the content of current object. There also are online help, save and exit functions. *DBE* is mouse operated and used offline.

R (Figure 5) allows for searching a databases using the AI^2R method.

Figure 5. Screen of Retrieval (*R*). The upper left scrolling window shows objects names of opened database to be searched. The upper right window shows result of search. The lower window shows content of current retrieved object. The narrow window in the middle serves to enter a new or load a previous query. There also are online help, save and exit functions. *R* can be used offline.

4. Conclusion

AI^2R was defined and its new mathematical properties were shown. An application was briefly described. Experience so far has confirmed expectation, i.e. high precision. Further application focuses on data fusion using AI^2R between several Internet searchers and user.

Acknowledgement

I thank G. Droszlei, J. Forman and Z. Papp for their help in implementing AI^2R.

References

[1] CARRICK, C. and WATTERS, C.: Automatic Association of News Items. *Information Processing and Management* **33(5)** (1997), 615-632.

[2] DOMINICH, S.: *The Formulation of the Interaction Information Retrieval Model As A New And Complementary Framework For Information Retrieval* Ph.D. Thesis, (in English), Hungarian Academy of Sciences, Budapest, Hungary, 1993.

[3] DOMINICH, S.: Interaction Information Retrieval. *Journal of Documentation* **50(3)** (1994), 197-212.

[4] DOMINICH, S.: The Interaction-Based Information Retrieval Paradigm. In: Kent, A. and Williams, J.G. (eds.) *Encyclopedia of Computer Science and Technology* Vol. 37, Suppl. 22, Marcel Dekker, Inc., New York Basel Hong Kong (1997a), 175-192.

[5] DOMINICH, S.: The Interaction-Based Information Retrieval Paradigm. In: Kent, A. (ed.) *Encyclopedia of Library and Information Science* Vol. 59, Suppl. 22, Marcel Dekker, Inc., New York Basel Hong Kong (1997b), 218-238.

[6] DOMINICH, S.: An I^2R (Interaction Information Retrieval) Pre-processor for Relevance Feedback. *Technology Letters* **2(1)** (1998), 5-18.

[7] KANG, H.K. and CHOI, K.S.: Two-level Document Ranking Using Mutual Information In Natural Language. *Information Processing and Management* **33(3)** (1997), 289-306.

[8] LIN, X.: Map Displays for Information Retrieval. *Journal of the American Society for Information Science* **48(1)** (1997), 40-54.

[9] LIU, G.Z.: Semantic Vector Space Model: Implementation and Evaluation. *Journal of the*

American Society for Information Science **48(5)** (1997), 395-417.

[10] MOCK, K.J. and VEMURI, V.R.: Information Filtering via Hill Climbing, Wordnet and Index Patterns. *Information Processing and Management* **33(5)** (1997), 633-644.

[11] PEARCE, C. and NICOLAS, C.: TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data. *Journal of the American Society for Information Science* **47(4)** (1996), 263-275.

[12] SALTON, G., ALLAN, J. and SINGHALL, A.: Automatic Text Decomposition and Structuring. *Information Processing and Management* **32(2)** (1996), 127-138.

[13] SALTON, G., SINGHALL, A., MITRA, M. and BUCKLEY, C.: Automatic Text Structuring and Summarization. *Information Processing and Management* **33(2)** (1997), 193-207.

[14] SHAW, W.M., BURGIU, R. and HOWELL, P.: Performance Standards and Evaluation In Information Retrieval Test Collection: Cluster-Based Retrieval Models. *Information Processing and Management*. **33(1)** (1997), 1-14.