Neuronové sítě a Information Retrieval

Tomáš Skopal*

Abstrakt

Tento přehledový článek popisuje několik oblastí Information Retrieval (resp. dokumentografických informačních systémů), do kterých pronikly neuronové sítě. Ve většině případů jde o aplikace, kde klasický analytický způsob zpracování selhává kvůli své sémantické složitosti nebo rozsáhlosti zkoumaných dat. Elegantním řešením v těchto situacích je použití sofistikovaných numerických metod, v našem případě neuronových sítí.

1 Přehled

- První příspěvek shrnuje použití neuronových sítí na poli Text Retrieval, tj. klasických dokumentografických informačních systémů. Viz kapitola 2.
- Předmětem druhého příspěvku je integrace heterogenních distribuovaných databází pomocí neuronových sítí. Ačkoliv se zde jedná o čistě databázovou problematiku, metody použité v procesu sémantické integrace jsou velmi specifické pro oblast Information Retrieval. Viz kapitola 3 a literatura [LCL00].

2 Tolerantní a adaptivní metody a Text Retrieval

Množství elektronicky dostupných informací se celosvětově rapidně zvětšuje. Většina z nich je tvořena klasickými textovými dokumenty, často přístupnými z internetu. Stává se čím dál těžším vyznat se v moři znalostí a informací a vybrat právě ty relevantní, které nás zajímají. Information Retrieval (IR) je disciplína, která popisuje metody hledání relevantních informací a z tohoto hlediska bude do budoucna jistě jednou z klíčových.

Současný stav IR vykazuje při hledání relevantních informací jisté nedostatky. Při rozsáhlých experimentech se potvrdilo, že tradiční metody IR označují za relevantní pouze zlomek dokumentů z kolekce všech skutečně (pro člověka) relevantních dokumentů (viz [VH98]). Proto jsou potřeba lepší modely pro IR. Uveďme hlavní nedostatky současných metod IR:

- Kognitivní modely se modelují striktně matematicky. Dotaz na dokument je vyhodnocován podobnostními funkcemi, které ne zcela odpovídají lidskému posuzování podobnosti. Vágnost, nezbytný nástroj lidského chápání a vyjadřování, není v současných IR systéch uspokojivě modelována.

*VŠB-Technical University Ostrava, Department of Computer Science

- Nedostatek adaptability. Matematické modely nerozlišují míru významnosti různých termů v rámci problémové domény a nejsou schopny se adaptovat na kombinace termů, tj. pracovat s kontextem.
- Tradiční IR systémy předpokládají homogenní datové zdroje, ačkoliv uživatel očekává heterogenní odpověď – odpověď složenou z dokumentů různých typů. Neřeší se problémy sémantického propojení multimediálních či vícejazyčných zdrojů.

Neuronové sítě v IR mohou některé uvedené aspekty odstranit nebo alespoň zmírnit.

2.1 Neuronové sítě v IR

Neuronové sítě zpracovávají data paralelně a distribuovaně a tím představují vysoce tolerantní a adaptivní systémy. Mohou se učit z existujících skutečností i v případě, kdy člověk nezná nebo není schopen určit skutečná pravidla modelované skutečnosti.

Paradigma *soft computing*, do kterého jistě patří také neuronové sítě, se zdá být ideální platformou pro modelování metod IR. V současné sobě existují čtyři kategorie Text Retrieval systémů využívajících neuronové sítě.

2.2 Kohonenovy mapy

Pro kategoriální klasifikaci velké kolekce dokumentů se používají Kohonenovy samoorganizované mapy (Kohonen Self-Organizing Maps (SOM)). Tato metoda provádí projekci z vysoce-rozměrného prostoru dokumentů do prostoru s nižší dimenzí – většinou dvourozměrné mřížky. Důležitou vlastností projekce je částečné zachování topologie, tj. blízké body v původním prostoru jsou blízko i v mřížce.

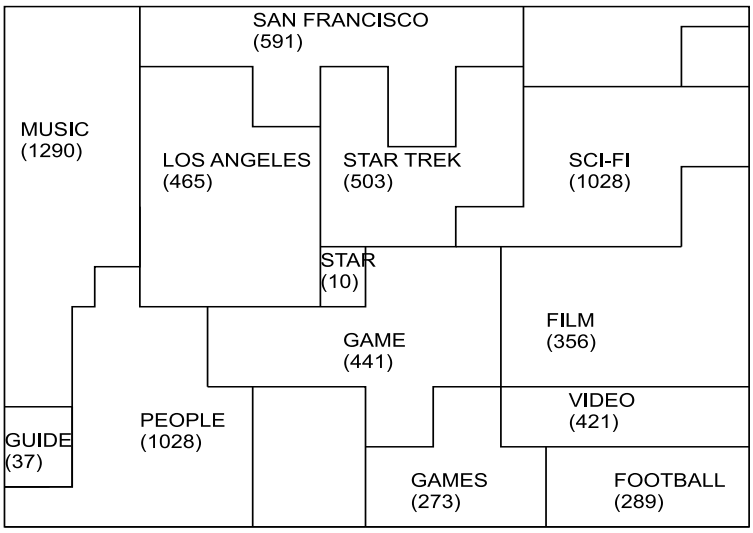
V případě IR, metoda SOM "vyrábí" *grafy podobnosti* (similarity graphs) mezi dokumenty. Na vstupu (ve vstupní vrstvě sítě) jsou vektory vlastností jednotlivých dokumentů. Metoda SOM v průběhu zpracování dokumentů vytváří v Kohonenově mapě shluky, ke kterým jsou přiřazovány zpracovávané dokumenty. Toto shlukování je prováděno na základě zvýšené excitace vítězného neuronu a jeho sousedů.

Dílčím problémem je velikost a reprezentace vektoru vlastností dokumentu, který vstupuje do neuronové sítě. V naivním vektorovém modelu je to například vážený histogram slov, která se v daném dokumentu nacházejí. Takováto reprezentace ale není vhodná, protože velikost vektoru (a tudíž i dimenze původního prostoru) je v řádu desetitisíců, což pro Kohonenovu mapu není z hlediska efektivity ideální. Pro tyto účely se používají různé předzpracující techniky redukce dimenze vektoru vlastností dokumentu:

- Latentně-sémantická indexace (LSI). Tato maticová metoda poměrně účinně redukuje dimenzi tak, že pomocí SVD (singular-value decomposition) rozkladu je vytvořen předem určený (malý) počet důležitých faktorů, které dohromady tvoří význam jednotlivých dokumentů. Do Kohonenovy mapy pak vstupují vektory s dimenzí rovnou počtu těchto faktorů.
- Náhodná projekce histogramů. Experimentálně se ukázalo, že prosté náhodné promítnutí vektoru histogramu do prostoru nižší dimenze vede k dobrým výsledkům bez ztráty signifikantních informací odlišujících dokumenty.

- Mapy pro slovní kategorie. Elegantním řešením je opět použití metody SOM pro redukci dimenze vstupního vektoru. Speciální "slovní" Kohonenova mapa nyní shlukuje slova do slovních kategorií, přičemž slučuje synonyma, různé tvary stejných slov a neslova do významových kategorií. Do původní "dokumentové" Kohonenovy mapy pak vstupují vektory s dimenzí rovnou počtu slovních kategorií.

Na obrázku 1 vidíme vytvořenou Kohonenovu mapu a shluky podobných dokumentů. Více o této problematice viz [Ko98].



Obrázek 1: Shluky dokumentů jako Kohonenova mapa

Dotaz do hotové mapy vrátí jako odpověď všechny dokumenty do určité vzdálenosti v nalezeném shluku.

2.3 Asociativní paměti

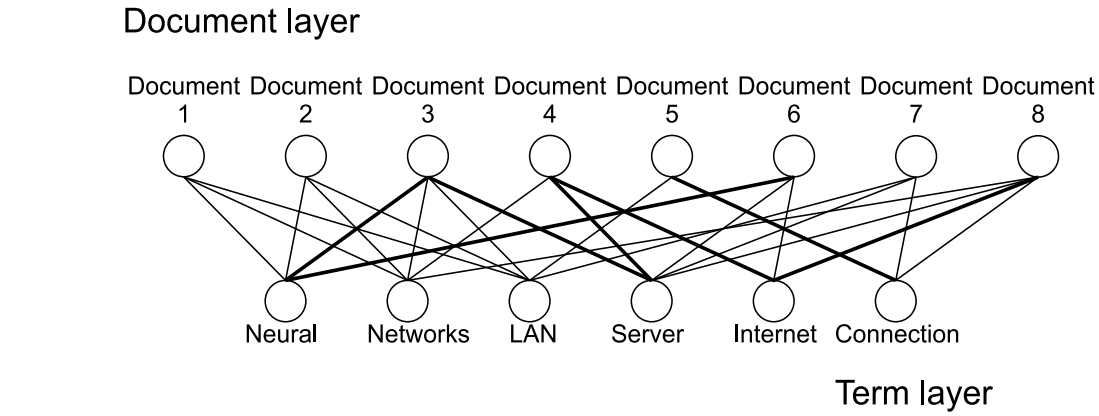
Asociativní paměti jako Hopfieldovy sítě jsou mocným nástrojem tolerujícím chyby. Dokumenty jsou uloženy jako energetická minima. Jako dotaz slouží zkreslený vzor. Síť minimalizuje jeho energii směrem k nejbližšímu minimu, které reprezentuje výsledný dokument sloužící jako odpověď. Bohužel, tato metoda vrací pouze jediný relevantní dokument jako odpověď.

Nicméně, asociativní paměti jsou v IR systémech úspěšně používány například v oblasti kontroly pravopisu.

2.4 Sítě s rozptýlenou aktivací

Nejběžnějšími a nejúspěšnějšími typy neuronových sítí v Information Retrieval jsou tzv. *sítě s rozptýlenou aktivací* (spreading activation networks) představené v [Be89] a [Kw89]. Jedná se o jednoduché dvouvrstvé neuronové sítě, kde vstupní vrstva reprezentuje termy a výstupní vrstva reprezentuje dokumenty. Váhy mezi oběma vrstvami jsou počátečně nastaveny podle výsledků vah tradičních indexovacích technik.

Uživatelský dotaz specifikovaný několika termy je zpracován tak, že vstupní neurony odpovídající dotazovaným termům jsou aktivovány a tato aktivace je *rozptýlena* do výstupní vrstvy a zpět. Nejvíce aktivované neurony výstupní vrstvy – dokumenty – jsou vráceny uživateli jako výsledek. Schéma vidíme na obrázku 2.



Obrázek 2: Síť s rozptýlenou aktivací. Tučné čáry představují vyšší váhy spojení neuronů.

Sítě s rozptýlenou aktivací prokázaly dobrou výkonnost v testech konferencí TREC, viz např. [BS98]. Bližší pohled na tyto sítě velmi připomíná klasický vektorový Information Retrieval model. V roce 1994 dokonce Mothe ukázal teoretický i empirický důkaz, že po prvním kroku aktivace neuronů jsou tyto metody takřka ekvivalentní.

Je vidět, že tento model dostatečně nevyužívá potenciálu neuronových sítí. Není možné jej trénovat, ani neobsahuje skryté vrstvy neuronů, které obyčejně zvyšují výpočetní schopnosti sítí.

2.5 Experimentální "Backpropagation" modely

Ačkoliv algoritmus backpropagation je jedním z nejúčinnějších nástrojů ze všech modelů neuronových sítí, nebyl do oblasti Information Retrieval příliš často aplikován.

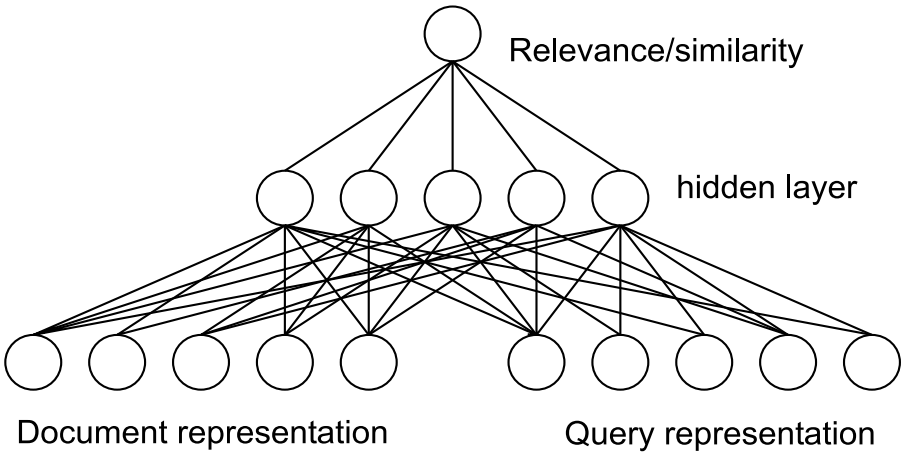
Model představený v [MCK90] je v podstatě rozšířením sítí s rozptýlenou aktivací o několik skrytých vrstev. Jako trénovací množina slouží množiny termů a k nim odpovídající množiny dokumentů. Ovšem zde je problém s výstupní – dokumentovou – vrstvou. Ta je pro trénovací proces příliš velká na to, aby se dala manuálně zvládnout.

2.6 Model COSIMIR

O více úspěšné zavedení metody Backpropagation do Information Retrieval se pokusil T. Mandl [Ma98, Ma00] se svým modelem COSIMIR (COgnitive SIMilarity learning in Information Retrieval). Důležitou podmínkou při návrhu bylo odstranění nedostatků předchozích metod založených na modelu Backpropagation, viz předchozí podkapitola. Klasické schéma – na vstupní vrstvě dotaz a na výtupní vrstvě dokumenty – bylo pozměněno. V modelu COSIMIR je obojí, jak dotaz, tak (jeden) dokument, na vstupu. Výstupní vrstva obsahuje

pouze jediný neuron, jehož excitace určuje relevanci, resp. podobnost vstupní dvojice dokument/dotaz. Skrytá vrstva propojuje obě vstupní vrstvy a skýtá další možnosti adaptace.

Na obrázku 3 je zobrazeno schéma modelu COSIMIR.



Obrázek 3: Model sítě COSIMIR

Obě dvě vstupní vrstvy jsou současně excitovány a aktivace je rozptýlena přes skrytou vrstvu do výstupního neuronu. Výsledná hodnota určuje podobnost/relevanci dotazu k dokumentu. Tento krok je opakován pro každý dokument v kolekci. Pomocí algoritmu Backpropagation formuje COSIMIR ve skryté vrstvě sub-symbolické reprezentace a implementuje tak složitou funkci podobnosti.

2.7 Shrnutí

Neuronové sítě jsou velmi tolerantní metody pro zpracování informací. Důsledkem toho subjektivní vyhodnocení jednoho uživatele a odlišné vyhodnocení jiného uživatele nevede k žádným dramatickým změnám ve výkonnosti sítě. Tradiční modely Information Retrieval používají matematické funkce podobnosti, například kosinus úhlu ve vektorovém modelu, pro určení relevance dokument/dotaz. Tyto jednoduché metody ovšem nepočítají se subjektivitou lidského rozhodování. Amos Tversky v roce 1977 ukázal, že lidsky chápaná podobnost nemusí být ani tranzitivní, dokonce ani symetrická (viz [Tv77]). Naproti tomu většina matematických funkcí podobnosti tyto podmínky splňuje.

Neuronové sítě, jako obecná numerická metoda, nemají žádný specifický matematický předpis, který by je omezoval. Spolu s tím, jak je síť trénována uživatelem, je do ní vnášena "lidská subjektivita". Důsledkem toho se může projevit, že některé termy jsou méně důležité než jiné, dokonce se může tímto způsobem modelovat kontextová závislost významu termů. Na druhé straně, jako daň za univerzálnost neuronových sítí, je potřeba sítě experimentálně a "heuristicky" vyladit tak, aby dávaly co nejlepší výsledky. Ani potom však není zaručena stoprocentní přesnost vyhodnocování. Ale s tímto rizikem se musí počítat v rámci celého paradigmatu *soft computing*.

3 Databázová integrace použitím neuronových sítí

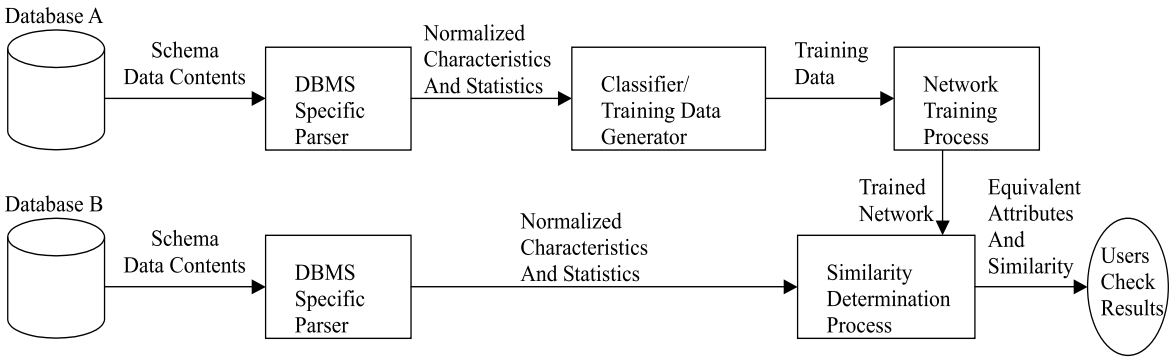
Aplikace v mnoha odvětvích průmyslu často vyžadují přístup do mnoha heterogenních distribuovaných databází. Klíčovým prvkem integrace heterogenních databází je sémantická integrace, tj. identifikace atributů z různých databází, které vyjadřují stejný koncept reálného světa. Ovšem pravidla pro tuto identifikaci nemohou být "předprogramována", protože míra korespondence jednoho atributu jedné databáze k jiným atributům jiných databází může být nejasná – fuzzy. Cílem je vytvořit integrační schémata, která podle nějakých kritérií významově shlukují různé atributy různých databází. Například atribut `Student_ID` z registrační databáze školy může (částečně) korespondovat (být ve shluku, resp. být podobný) s atributem `SSN` (Social Security Number) v databázi pojišťovací společnosti. Je zřejmé, že manuální porovnání všech možných n-tic atributů všech databází je prakticky neuskutečnitelná věc.

Pro představu, společnost US West spravuje 5 TB dat organizovaných v tisíci databázích, kde samotná informace o zákazníkovi je rozptýlena přes 200 různých databází. Jiným příkladem může být dílčí databáze firmy Boeing Computer Services, která má stovky atributů a miliony záznamů.

V článku [LCL00] navrhli Li, Clifton a Liu postup založený na vytvoření integračního schématu pomocí neuronových sítí.

3.1 Stručně sémantická integrace

Princip této metody je založen na vygenerování atributových shluků referenční (můžeme také říct trénovací) databáze, kde pro každý zkoumaný atribut z referenční databáze je přiřazen stupeň fuzzy příslušnosti ke každému shluku. Tím je integrační schéma vytvořeno. Ve skutečnosti je tímto schématem naučená neuronová síť. V druhé fázi (Similarity Determination) se integračnímu schématu (síti) předkládají různé atributy z různých databází a těm je přiřazena fuzzy příslušnost ke všem atributovým shlukům. Ve třetí fázi uživatel subjektivně kontroluje správnost shlukování.



Obrázek 4: Přehled metody sémantické integrace

Na obrázku 4 vidíme schématicky celý proces sémantické integrace. Nejprve se z databáze extrahují vlastnosti jednotlivých atributů (na obrázku proces *DBMS Specific Parser*). Z vlastností atributů a metadat referenční databáze A jsou následně vygenerovány shluky po-

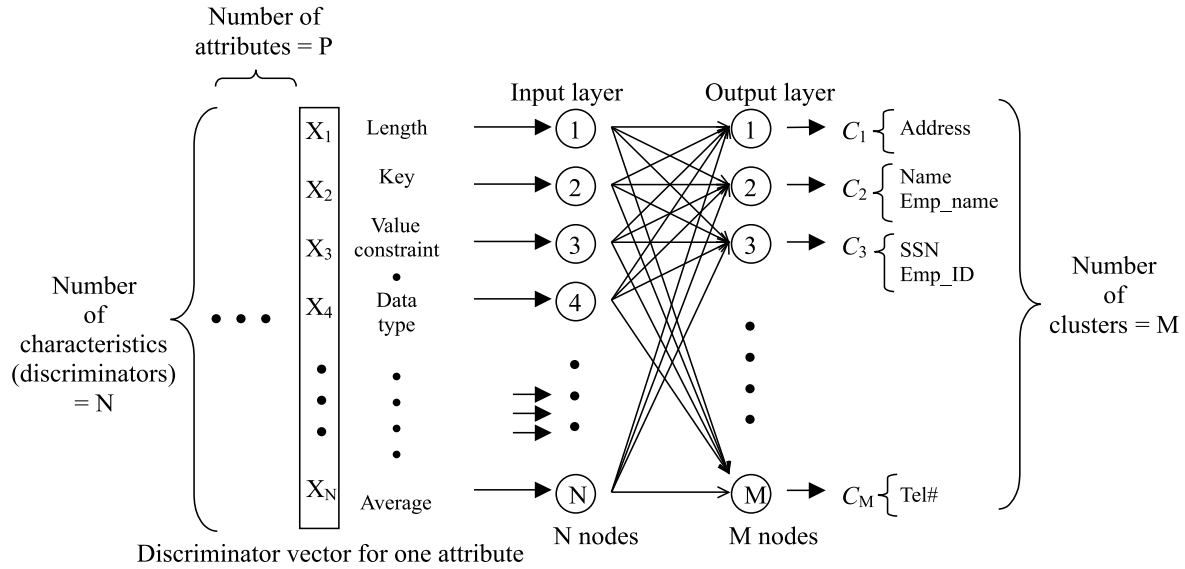
dobných (ve smyslu extrahovaných vlastností) atributů (na obrázku proces *Classifier/Training Data Generator*). Tyto shluky jsou předány neuronové síti jako požadovaný výstup trénovací množiny. Neuronové síti se pak předkládají trénovací vzory (atributy referenční databáze) a klasickou metodou back-propagation jsou upravovány parametry sítě tak, aby síť přiřadila atribut ke správnému shluku – samozřejmě s nějakou prahovou příslušností (na obrázku proces *Network Training Process*). Na naučené síti jsou pak testovány atributy ostatních databází a jsou jim přiřazovány stupně příslušnosti ke všem shlukům (na obrázku proces *Similarity Determination Process*). V závěrečné fázi uživatel kontroluje relevaci shluků a jim přiřazených atributů z ostatních databází.

3.2 Extrakce atributových vlastností

Každý atribut databáze má vlastnosti jako `Length` – délka atributu v bytech, `Key` – binární vlastnost, zda atribut je klíčem či nikoliv, `Value constraint` – omezení atributu, `Data type` – datový typ, a další vlastnosti, které mohou být zjištěny ze schémat a metadat příslušné databáze. Všechny tyto vlastnosti atributu jsou následně normalizovány do numerických hodnot. Získaný vektor vlastností atributu nazýváme *discriminator vector*. Každý atribut reprezentovaný vektorem o N složkách si můžeme představit jako bod v N-rozměrném prostoru. Neuronová síť pak nedělá nic jiného, než že vytváří v tomto prostoru shluky blízkých bodů – tj. podobných atributů.

3.3 Klasifikace shluků

Pro počáteční klasifikaci shluků byla použita samoorganizovaná dvouvrstvá neuronová síť – metoda Self-Organizing Map (SOM), představená Kohonenem v roce 1987. Uživatel zadá

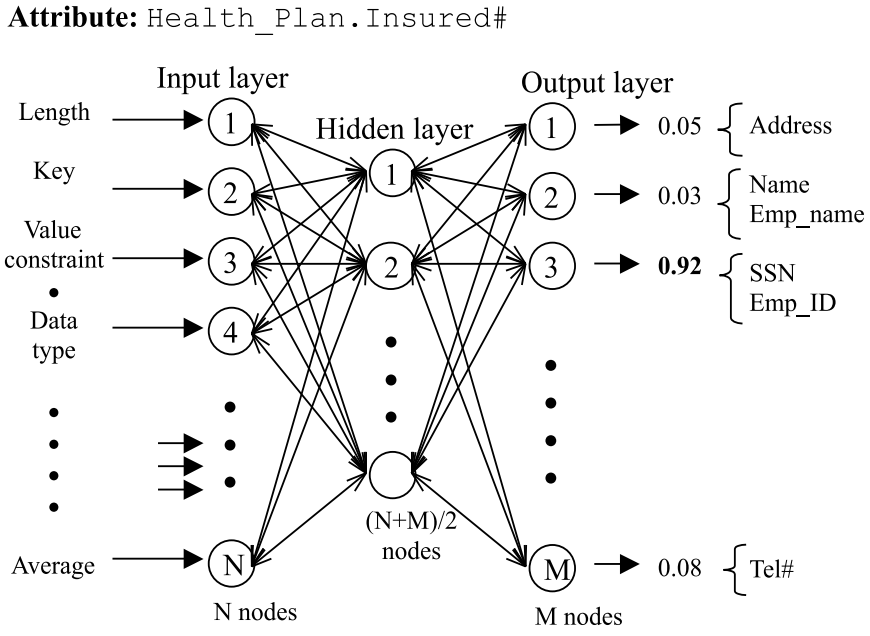


Obrázek 5: Síťová architektura Self-Organizing Map

pouze počet shluků M, které mají být vytvořeny – ty určují počet M neuronů výstupní vrstvy. Vstupem je P vektorů atributových vlastností (discriminators) referenční databáze o velikosti vektoru N – tato velikost určuje počet N neuronů vstupní vrstvy. Metoda SOM následně vytvoří požadovaný počet shluků atributů. Celou situaci vidíme na obrázku 5.

3.4 Učení neuronové sítě

Nejdůležitější část procesu sémantické integrace je učení a použití třívrstvé neuronové sítě pro rozpoznávání kategorií atributů a jejich přiřazení k příslušným shlukům. Vstupem je trénovací množina vektorů vlastností referenční databázem, tj. vstupní vrstva sítě má opět N neuronů. Stejně tak jako v předchozím případě, výstupní vrstva je tvořena M neurony, které představují jednotlivé shluky určené v předchozí fázi. Mimo to, existuje zde prostřední (skrytá) vrstva o $(N+M)/2$ neuronech (viz obrázek 6). Tento počet byl zvolen s ohledem na výkonnost učení sítě při pozdějších experimentech. Síť je trénována klasickou back-propagation metodou, uvedenou Rumehartem v roce 1986.



Obrázek 6: Učení a použití neuronové sítě

3.5 Použití neuronové sítě

V této fázi je síť připravena k použití na ostatních databázích. Vektory jednotlivých atributů jsou přiřazovány ke shlukům.

Na obrázku 6 vidíme, že atribut `Health_Plan.Insured#` byl přiřazen s vysokým stupněm příslušnosti ke shluku {`Emp_ID, SSN`} – půjde tedy zřejmě o nějaký identifikátor stejného typu

objektu (zde osoby). Důsledkem tohoto shluku mohou být například při integraci databází sjednoceny závislosti na těchto atributech tak, jako by to byl jediný atribut.

3.6 Shrnutí

Metoda sémantické integrace databází může být užitečným nástrojem při automatizovaném generování společných databázových schémat v oblasti rozsáhlých heterogenních distribuovaných databází.

Podrobnější popis metody a zejména výsledky experimentů jsou k dispozici v literatuře [LCL00].

Reference

[Be89] R.Belew, *Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents.* In: Belkin and Rijsbergen 1989. pp. 11-20.

[BS98] M.Boughanem, C.Soulé-Dupuy, *Mercure at trec6.* In: Voorhees and Harman 1998.

[Ko98] T.Kohonen, *Self-organization of very large document collections: State of the art.* Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, volume 1, pages 65-74. Springer, London, 1998

[Kw89] K. L. Kwok, *A Neural Network for Probabilistic Information Retrieval.* In: Belkin and Rijsbergen 1989. pp. 21-30.

[LCL00] W.Li, C.Clifton, S.Liu. *Database Integration Using Neural Networks: Implementation and Experiences*, Knowledge and Information Systems, Springer-Verlag London Ltd., 2000

[Ma98] T.Mandl, *Das COSIMIR Modell: Information Retrieval mit dem Backpropagation Algorithmus.* ELVIRA-Arbeitsbericht 10, IZ Sozialwissenschaften, Bonn, 1998.

[Ma00] T.Mandl *Tolerant and Adaptive Information Retrieval with Neural Networks.* In: Global Dialogue. Science and Technology Thinking the Future at EXPO 2000 Hannover. 11.-13.7.2000.

[MCK90] H.Mori, C.Chung, Y.Kinoe, Y.Hayashi. *An Adaptive Document Retrieval System Using a Neural Network.* In: International Journal of Human-Computer Interaction 2 (3). pp. 267-280., 1990

[Tv77] A.Tversky, *Features of Similarity.* In: Psychological Review vol. 84 (4). pp. 327, 1977

[VH98] Voorhees E, Harman D (eds.) (1998). The Sixth Text Retrieval Conference (TREC-6). NIST Special Publication 500-240. National Institute of Standards and Technology. Gaithersburg. Nov. 19-21 1996.