

Włodzisław Duch
Uniwersytet Mikołaja Kopernika w Toruniu

Publikacje naukowe o sieciach neuronowych w Internecie

Młode pokolenie studentów i naukowców unika chodzenia do bibliotek, ponieważ uważają oni, że wszystko jest do odszukania w Internecie, a jeśli nie są w stanie wyłowić stamtąd żadnej informacji, to znaczy, że ona w ogóle nie istnieje. W zeszłym roku student pewnej uczelni amerykańskiej przysłał mi emaila z zapytaniem: "kim był ten facet Kopernik? Mam napisać o nim pracę i nic nie mogę znaleźć....."

Dla młodego pokolenia Internet to nie jest zwykłe źródło informacji, ale **jedynę** jej **źródło**. Pierwszą rzeczą, jaką robią młodzi pracownicy naukowcy, aby odszukać potrzebne im artykuły, to przeszukanie stron internetowych czasopism. Dopiero później - jeśli mają szczęście i trafią na dobrą bibliotekę prenumerującą to właśnie czasopismo - idą do biblioteki. Niestety, nawet w Stanach czy też Europie Zachodniej, nie mówiąc już o krajach środkowoeuropejskich oraz rozwijających się, trudno znaleźć dobrą bibliotekę. Wiele osób (łącznie ze mną) przerzuciło się z inżynierii, fizyki czy też matematyki na neuronową technikę komputerową. Fakt, że mogli oni zgromadzić wiele przydatnych im artykułów, kopiując je z archiwów internetowych, w znacznym stopniu ułatwił ten krok. Jednak wielu "poważnych" naukowców lekceważy nowe media i przeznaczają znaczną część czasu na przygotowywanie swoich artykułów i wystąpień na konferencje, a bardzo niewiele czasu na tworzenie archiwów ze swoimi pracami oraz stron WWW, prezentujących ich opracowania. Przyznaje się różne nagrody za najlepsze artykuły, ale nie docenia się najbardziej użytecznych stron WWW, dlatego też brak jest motywacji do ich tworzenia, bądź aktualizowania.

Niektóre materiały konferencyjne publikowane są w grubych tomach, mają tysiące stron. Kto jest w stanie to wszystko przeczytać? Szybciej jest odszukać odpowiedni referat w Internecie i wydrukować tylko te materiały, które są potrzebne. Środowiska osób zainteresowanych sieciami neuronowymi (INNS, ENNS, APNNA, IEEE NNC), redakcje czasopism i organizatorzy większych konferencji w dalszym ciągu funkcjonują przeważnie "po staremu": nie publikując adresów WWW uczestników konferencji, często nie podając nawet adresów poczty elektronicznej,, a na stronach domowych konferencji rzadko można znaleźć spisy treści sprawozdań. Są pewne oznaki zmian, przynajmniej jeśli chodzi o konferencje IEEE: wspaniały nowy serwis IEEE Xplore (była wzmianka o nim w artykule redakcyjnym IEEE TNN ze stycznia 2000 roku) umożliwia każdemu dostęp do spisów treści IEEE Transactions, czasopism, sprawozdań konferencyjnych oraz norm. Członkowie IEEE mają także pełen dostęp do artykułów i innych materiałów w formacie PDF. Publikacje te powinny być powiązane (cross-linked) za pomocą hiperłączy. Jest to przydatne uzupełnienie bibliografii IEEE on-line oraz serwisów Opera. IEEE Xplore jest dostępny pod adresem: <http://ieeexplore.ieee.org>.

Wielu specjalistów prowadzi swoje lokalne archiwa, jednak bez dobrze zorganizowanej centralnej bazy skatalogowanych materiałów nie tylko trudno je odnaleźć, autorzy nie mają nawet silnej motywacji do umieszczania swoich prac w lokalnych archiwach. Bez wsparcia instytucji funkcjonujących w sieciach neuronowych wiele pożytecznych inicjatyw, dążących do tworzenia bazy materiałów naukowych, umiera śmiercią naturalną. Na przykład archiwum Neuroprose, które utworzono w 1989 roku, zawierające ponad 600 referatów i prac doktorskich, nigdy nie było należycie prowadzone. Referaty automatycznie "wrzucano" do olbrzimego katalogu, który tak się rozrósł, że nie sposób było znaleźć w nim coś ciekawego, chyba że znało się nazwę pliku. Ostatni raz modyfikowano plik "czytaj to" w 1994 roku, a archiwum zdaje się przestało istnieć w roku 1998. Neuroprose opierał się na protokole ftp, który jest już przestarzały i często nie działa prawidłowo, co spowodowane jest różnymi firewallami, zainstalowanymi ze względów bezpieczeństwa. Umieszczanie plików w jednym, dużym katalogu nie jest na dłuższą metę dobrym rozwiązaniem. Nie ma powodu, dla którego archiwa nie mogłyby być dostępne poprzez strony WWW. London and South-East Center for High Performance Computing (SEL-HPC) uruchomiło archiwum, w którym przechowywane są publikacje dotyczące programowania równoległego i funkcjonalnego, przetwarzania obrazów, matematyki komputerowej, sieci neuronowych, relacji człowiek-komputer oraz innych tematów. W archiwum tym zgromadzono ponad 7000 publikacji z linkami do stron domowych osób, których prace były tam przechowywane. Niestety w 1998 roku Centrum SEL-HPC przestało funkcjonować i chociaż archiwum nadal istnieje, to nie są do niego włączane nowe prace, pojawia się jedynie komunikat: "wzbroniony dostęp".

Internet przepełniony jest starymi plikami, których nie usunięto i które pojawiają się przy każdym wyszukiwaniu. Wystarczy spytać o "archiwa neuronowe", a dojdzie się do jakiegoś niedziałającego już linku lub wykazu baz archiwalnych z 1993 roku, które od lat nie istnieją. Niewielu ludzi jest na tyle odpowiedzialnych, aby usuwać przedawnione pliki i wpisywać datę ostatniej modyfikacji swojej strony. W dalszej części artykułu przedstawiam krótki przegląd aktualnej sytuacji. Dogłębna ocena tych oraz innych przydatnych projektów będzie mile widziana.

Los Alamos e-Print Archive (<http://arxiv.org/>) nieprzerwanie od sierpnia 1991 roku służy fizykom i matematykom. Codziennie łączy się z tym archiwum 9000 hostów (według danych z kwietnia 2000 roku), wyłączając z tego liczne kopie archiwum (mirror sites). Co miesiąc, przybywa ponad 2500 nowych publikacji. Obecnie szacuje się, że w archiwum tym znajduje się ogółem około 130.000 prac. Są tam też pewne interesujące dla zajmujących się sieciami neuronowymi podkatalogi, takie jak:

- **Computing Research Repository, CoRR** (<http://arxiv.org/archive/cs/intro.html>) - uruchomione we wrześniu 1998 roku, sponsorowane przez ACM. Publikacje zamieszczane w CoRR są klasyfikowane na dwa sposoby: według tematyki (każdy temat ma moderatora) oraz z zastosowaniem układu klasyfikacyjnego ACM, który dotyczy całej informatyki. Zagadnienia związane z inteligencją elektroniczną, to: sztuczna inteligencja, informatyka a język, obrazy komputerowe i rozpoznawanie znaków, relacja człowiek-komputer, uczenie się, systemy wieloczynnikowe, modelowanie sieci neuronowych i ewolucyjna technika komputerowa oraz robotyka.
- **Nonlinear Sciences.** Archiwum zawierające informacje o: adaptacyjnych i samoorganizujących się systemach, automatach komórkowych i dynamice chaotycznej.
- **Physics.** W archiwum tym znajdują się materiały o: systemach nieuporządkowanych i sieciach neuronowych, analizie danych, statystyce, prawdopodobieństwie oraz analizie Bayesa.
- Inne ciekawe bazy tego archiwum, takie jak nauki na poziomie komórek czy molekuł, neuronauki rozwojowe i behawioralne/systemowe, nie są jeszcze zbyt popularne, ale mogą być w nich umieszczane nowe publikacje.

E-Print Archive ma wiele wersji na całym świecie i jest to bardzo dobra baza dla fizyków i matematyków. Podobną inicjatywę podjęto w dziedzinie nauk kognitywnych (nauk o poznawaniu, cognitive science). **CogPrints Archive** (<http://cogprints.soton.ac.uk/>) będzie gromadzić publikacje z dziedziny psychologii, nauki o neuronach, lingwistyki, sztucznej inteligencji, robotyki, widzenia, uczenia się, mowy, sieci neuronowych, filozofii umysłu i mowy, ekologii behawioralnej, socjobiologii, genetyki behawioralnej, teorii ewolucji, psychiatrii, neurologii, genetyki człowieka, skanowania mózgu, antropologii oraz innych nauk społecznych i matematycznych, odnoszących się do nauki o poznaniu. Archiwum CogPrints otrzymało nagrodę Psychological Science Award od PsychologicalScience.net za "wkład w psychologię w Internecie".

Prawdopodobnie nie ma żadnych, podobnych baz dla naukowców zajmujących się sieciami neuronowymi, mimo to warto jednak wspomnieć o kilku inicjatywach. Zamierzeniem ich jest indeksowanie zasobów sieci WWW, poszukiwanie prac w różnych archiwach prywatnych. Przechowywanie linków do publikacji naukowych na różnych stronach ma pewne wady: czasami linki są zmieniane lub też administratorzy systemów usuwają niektóre publikacje. Trudno jest indeksować materiały tak, aby były one przydatne. Z drugiej strony o wiele łatwiej jest indeksować różne publikacje w sieci WWW, niż tworzyć jakąś centralną bazę.

Zespół Princeton NEC Research Institute (Lee Giles, Steve Lawrence i Kurt Bollacker) stworzył **CiteSeer**, który obecnie nazywa się **Research Index** (<http://www.researchindex.com>). Jest to system służący do automatycznego tworzenia bibliotek cyfrowych, szczególnie nastawiony na znajdowanie cytowań, indeksowanie i ustalanie popularności prac. W wyniku prowadzonych poszukiwań otrzymuje się linki do dokumentów postscriptowych lub PDF i udostępniane są artykuły z oryginalnych źródeł lub lokalnych kopii; opcja "korekta" pozwoli wyszukać pierwszą stronę publikacji i wyświetli ją w przeglądarce. Autorzy powinni sprawdzać poprawność i nanosić poprawki w pisowni imion i nazwisk, szczególnie w przypadku stosowania znaków diakrytycznych - system ten nie radzi sobie z francuskimi, hiszpańskimi czy też polskimi nazwiskami! Na żądanie można obejrzeć cytowania w takim kontekście w jakim się one pojawiły. Zdaje się, że bardzo popularny jest Research Index, ponieważ dużo osób łączy się z nim nawet w środku nocy (czasu Princeton). "Computer Science Directory" podaje wykaz publikacji z różnych dziedzin, który jest uszeregowany według ilości cytowań, rangi autorów oraz wartości naukowej tych prac.

Obecnie Research Index jest z pewnością najlepszą bazą bibliograficzną. Nie ma tam co prawda odsyłaczy do stron domowych autorów (trudno by to było zrobić w sposób automatyczny, ponieważ autorzy nie podają w publikacjach adresów swoich stron WWW), ale często można do nich dojść rozszyfrowując link do ich publikacji. W bazie **HP-search** (Trier University, Niemcy, <http://hpsearch.uni-trier.de/hp/>) najłatwiej jest znaleźć strony domowe informatyków. W kwietniu 2000 roku znajdowało się tam ponad 42.000 opisów.

Collection of Computer Science Bibliographies (<http://liinwww.ira.uka.de/bibliography/index.html>) ma ponad milion odsyłaczy, co oznacza 660 MB opisów BibTeX! Jest to metabaza, która zawiera około 1200 specjalistycznych bibliografii i jest co miesiąc aktualizowana na podstawie ich oryginalnych wersji. Około 90.000 odsyłaczy zawiera adresy URL dokumentów online. Znaleźć tam można ponad 2000 linków do innych stron z informacją bibliograficzną, także do wielkiej Computer Science Bibliography w Trier University, gdzie przechowywane są informacje

bibliograficzne, dotyczące ważniejszych konferencji, książek oraz czasopism, ale jest tam stosunkowo niewiele informacji o sieciach neuronowych.

Wyszukiwarka **ML Papers** (<http://gubbio.cs.berkeley.edu/mlpapers/>) do indeksowania publikacji machine learning, która funkcjonuje od 1997 roku, była prawdopodobnie pierwszym narzędziem wprowadzającym automatyczne wybieranie tytułów, autorów i abstraktów z postscriptowych wersji publikacji. Interfejs do wyszukiwania jest prosty. Wyświetlane są tytuły, nazwiska autorów, abstrakty oraz linki do dokumentów postscriptowych (zdaje się, że jest to jedyny indeksowany format). W kwietniu 2000 roku można było znaleźć tam prawie 1300 prac ze słowem kluczowym "neural". Aby włączyć nowe prace wystarczy tylko podać adres URL do archiwum FTP lub HTTP.

Firma Just Research, która opracowała specjalną wyszukiwarkę o nazwie **Cora** (<http://www.cora.justresearch.com/>), obejmującą ponad 50.000 prac naukowych z około stu wydziałów informatycznych, podeszła do problemu automatycznego indeksowania w sposób interesujący. Informatyka została podzielona na 10 kategorii z tematami pod hasłem "Sztuczna Inteligencja", uwzględniającymi takie zagadnienia jak szukanie wiedzy w danych i uczenie maszynowe, a pod hasłem "Machine Learning" uwzględnione są sieci neuronowe, a także uczenie oparte na precedensach, algorytmy genetyczne, metody probabilistyczne, uczenie z krytykiem, indukcja reguł i zagadnienia teoretyczne. W bazie znajduje się 75 kategorii z dziedziny informatyki. Cora umożliwia wyszukiwanie słów kluczowych, które występują w publikacjach przechowywanych w formacie postscriptowym. Wyniki wyszukiwania zapewniają automatycznie generowane opisy BibTeX (czasami z błędami), zawierające: tytuł, autorów, abstrakty prac, adres strony głównej, na której znajdowała się dana praca (również alternatywne miejsca jej przechowywania), wykaz odsyłaczy wziętych z pracy i odsyłacze wsteczne (publikacje odnoszące się do tej właśnie pracy).

Dodawanie nowych plików jest bardzo proste, polega jedynie na podaniu adresu URL archiwum. Automatycznie wyświetla się 50 najwyższej cenionych prac z każdej kategorii, a oceniane są one według analizy cytowań, która pozwala automatycznie określić czy dana praca jest artykułem krytycznym, odkrywczym, czy ma bogatą bibliografię i czy została napisana przez autora o wysokiej randze. Do automatycznej klasyfikacji użyto uczenia z krytykiem i technik probabilistycznych. Projektem Cora kieruje Andrew McCallum. Pomimo, że przedsięwzięcie to jest jeszcze na etapie badań, jest ono całkiem przydatne.

W ramach projektu **New Zealand Digital Library, NZDL** (<http://www.nzdl.org/>) stworzono ogólnodostępne narzędzia wyszukujące, obejmujące szeroki zakres tematów, od raportów technicznych z dziedziny informatyki do filmów muzycznych. W projekcie szczególnie skoncentrowano się na tworzeniu bibliotek cyfrowych, posiadających funkcję przeszukiwania pełnotekstowego; w zautomatyzowanym tworzeniu wyszukiwarek nie stosuje się technologii komputerowego uczenia się. Jest tam bardzo duży indeks obejmujący ponad 45.000 raportów z dziedziny informatyki. Analizowano całe prace (ponad 30GB plików postscriptowych, 1.3 mln stron), tak więc można odszukać każdy fragment tekstu. W bazie znajduje się też 30.000 danych liczbowych wydobytych z raportów.

Networked Computer Science Technical Reference Library (NCSTRL) postawiła sobie za cel stworzenie "najważniejszego międzynarodowego magazynu raportów technicznych on-line z dziedziny informatyki". Dzięki udostępnianemu oprogramowaniu, różne instytucje (jest ich teraz ponad 160, w tym przodujące uniwersytety amerykańskie) mogą tworzyć biblioteki cyfrowe, które zostają później włączane do NCSTRL. Możliwe jest też prostsze rozwiązanie wykorzystujące serwery FTP, wymagające dostarczenia informacji bibliograficznych do centralnego indeksu NCSTRL (<http://cs-tr.cs.cornell.edu/>). Ponieważ indeksowanie nie jest automatyczne, projekt ten znajdzie większe zainteresowanie w środowiskach bibliotekarskich (funkcjonuje on w Cornell University Digital Library Research Group) niż wśród informatyków.

Istnieją eksperymentalne metody wyszukiwania informacji w oparciu o samoorganizujące się mapy (na przykład WebSOM, <http://websom.hut.fi/websom/>), ale jak do tej pory nie są one przydatne do większych wyszukiwań.

Na koniec **IEEE Bibliographies Online** (tylko dla członków IEEE), która dostarcza informacje bibliograficzne, dotyczące konferencji i czasopism sponsorowanych przez IEEE, ale bez linków do samych prac (jest to jedynie dostępne dla prenumeratorów za pośrednictwem serwisu Opera). Bibliografie IEEE można znaleźć pod adresem: http://www.biblio.ieee.org/scripts/biblio_home.html

Linki do wszystkich wymienionych tu serwisów (a także wielu innych) są gromadzone pod adresem: <http://www.phys.uni.torun.pl/~duch/neural.html#biblio>

Pierwotny adres: <http://ebib.oss.wroc.pl/2001/19/duch.html>



