Spis treści

O autorkach

Dr Ranjana Rajnish jest adiunktem na Wydziale Technologii Informacyjnej w Amity University w indyjskim Lucknow w stanie Uttar Pradesh. Dr Ranjana posiada ponad 25-letnie doświadczenie akademickie/badawcze. Współpracowała z takimi instytucjami, jak UP Technical University i Amity University, gdzie zajmowała stanowiska od wykładowcy informatyki po kierownika merytorycznego. Uzyskała stopień doktora informatyki, a jej obszar badań i nauczania obejmuje języki programowania, takie jak C, Python czy powłoka Borne Shell, a także inżynierię oprogramowania, eksplorację opinii/analizę nastrojów oraz opiekę zdrowotną. Na swoim koncie ma ponad 40 publikacji prezentowanych na renomowanych konferencjach krajowych i zagranicznych.

Dr Meenakshi Srivastava od 2005 r. pracuje jako adiunkt na Wydziale Technologii Informacyjnej w Amity University w indyjskim Lucknow, Uttar Pradesh. Uzyskała stopień doktora nauk inżynieryjnych. Jej obszar badań i nauczania obejmuje web mining, przetwarzanie obrazów, analizę biznesową i bioinformatykę. Ma ponad 18-letnie doświadczenie w nauczaniu i uczyła różnych języków programowania, m.in. C, C++, C#, JAVA, Python i R. Na swoim koncie ma ponad 40 krajowych i międzynarodowych publikacji.

O recenzencie

Dr Deepak Singh uzyskał stopień doktora w Indyjskim Instytucie Technologii w Kanpur i posiada 13-letnie doświadczenie w środowisku akademickim i branży IT. Specjalizuje się głównie w rzeczywistych zastosowaniach sztucznej inteligencji i systemów opartych na Internecie rzeczy. Oprócz nauczania w renomowanych instytucjach w Lucknow, świadczy również usługi konsultingowe dla kilku firm IT. Doradzał również władzom Uttar Pradesh w zakresie wdrażania rozwiązań informatycznych w bibliotekach publicznych.

Odegrał kluczową rolę w kilku projektach badawczych opartych na sztucznej inteligencji, w tym również w transformacji cyfrowej kilku czołowych instytutów edukacyjnych w Lucknow, dwóch projektach finansowanych przez rząd Indii oraz kilku innych projektach związanych z wdrażaniem rozwiązań sztucznej inteligencji i uczenia się maszyn. Prowadził szkolenia dla menedżerów wyższego szczebla w firmach IT w różnych częściach świata, m.in. w Singapurze, Japonii i Chinach.

Podziękowania

Podziękowania są najpiękniejszą częścią każdego projektu, ponieważ dają one możliwość wyrażenia wdzięczności osobom, które pomagały lub zapewniały ciągłe wsparcie w trakcie realizacji projektu.

Pisanie książki samo w sobie jest projektem. W czasie podróży związanej z tworzeniem tej książki, w mniejszym lub większym stopniu udzieliło nam pomocy wiele osób.

Nasza lista podziękowań zaczyna się od wszechmocnego Boga.

Jesteśmy wdzięczni pani Prernie Mishrze, która miała swój udział w powstaniu rozdziału 5, a także naszym studentom, panu Ayushowi Kumarowi Rathoremu, panu Anantowi Gupcie oraz panu Apurvie, którzy pomogli w wykonaniu kilku skryptów i rysunków do książki.

Dziękujemy dr Deepakowi Singhowi, którego cenne uwagi pomogły nam ulepszyć treść tej książki. Wspierał nas w całym procesie jej powstawania.

Wyrażamy również wdzięczność zespołowi BPB Publications za ich ciągłe wsparcie i pomoc w ukończeniu tej książki oraz umożliwienie jej publikacji. Jesteśmy również wdzięczni innym redaktorom w BPB, których cenny wkład pomógł nam w napisaniu tej książki.

Na koniec chcielibyśmy podziękować naszym rodzinom, a zwłaszcza naszym dzieciom. Bez ich wsparcia ukończenie tej książki nie byłoby możliwe.

Przyjemnej nauki.

Przedmowa

Pomysł na stworzenie tej książki pojawił się podczas prowadzenia badań z wykorzystaniem narzędzi do eksploracji danych w sieci Web (Web Data Mining). Miałyśmy spory problem ze znalezieniem jednej książki, która pomogłaby nam zrozumieć podstawy tego zagadnienia wraz z podaniem jego przykładowych implementacji. Ponadto większość autorów książek zakładała, że czytelnicy mają już wcześniejszą wiedzę na temat języka Python. W tej książce temat eksploracji sieci Web i języka Python prezentujemy od podstaw, tak aby był on zrozumiały nawet dla początkujących osób.

Zanim zaprosimy czytelników do zapoznania się z treścią tej książki, chciałybyśmy skorzystać z okazji i w ramach tej przedmowy przedstawić czytelnikom jej ogólny zarys.

Obecnie bardzo często zdarza się, że ilekroć przychodzi nam do głowy jakieś pytanie, zazwyczaj wyszukujemy je w Internecie za pomocą dowolnej wyszukiwarki (takiej jak Google, Yahoo itd.). Przy ponad 560 milionach użytkowników Internetu, Indie zajmują drugie miejsce pod względem jego wykorzystania, plasując się tuż za Chinami. To pozwala nam lepiej zrozumieć liczbę osób, jaka uzyskuje dostęp do Internetu w różnych celach, takich jak handel elektroniczny, media społecznościowe, e-learning itd.

Przedsiębiorstwa coraz szybciej zmierzają w kierunku monetyzacji zasobów cyfrowych.

Handel elektroniczny narodził się w maju 1989 r., kiedy to firma Sequoia Data wprowadziła na rynek Compumarket – pierwszy internetowy system dla handlu elektronicznego. Ten system wspierał sprzedawców w wystawianiu przedmiotów na sprzedaż i umożliwiał klientom przeszukiwanie bazy danych pod kątem interesujących przedmiotów i dokonywanie zakupów. Compumarket pozwalał również na korzystanie z kart kredytowych. Od tamtej pory handel elektroniczny zmienił globalną sprzedaż detaliczną, a organizacje biznesowe zaczęły domagać się zautomatyzowanych inteligentnych rozwiązań, takich jak przewidywanie oczekiwań i wymagań klientów. Te informacje nie tylko przekładają się na podniesienie poziomu satysfakcji klienta, ale także pomagają sprzedawcom

w utrzymywaniu odpowiednich zapasów magazynowych i zapewniają dane pozwalające na zwiększenie wielkości sprzedaży. Web mining odgrywa kluczową rolę w przewidywaniu takich oczekiwań klientów i wymaganych zapasów. Te informacje mogą być wykorzystywane przez firmy do efektywnego podejmowania decyzji. W podobny sposób analizowane mogą być dane z różnych serwisów społecznościowych, które mogą być następnie wykorzystywane przez niektóre branże na potrzeby rozwoju biznesu, badań społecznych, usług zdrowotnych czy edukacji. Szerokie możliwości zastosowania eksploracji sieci Web stanowiły główną motywację do napisania tej książki.

My jako autorki chciałyśmy mieć kompleksowe rozwiązanie do tworzenia aplikacji opartych na eksploracji sieci Web. Ta książka stanowi pełny przegląd w zakresie teoretycznego i matematycznego modelowania różnych algorytmów używanych do eksplorowania sieci, a także rozwiązań opartych na języku Python do ich implementacji.

Proponowana książka *Web Data Mining z użyciem języka Python* opiera się na koncepcji i zastosowaniach eksploracji informacji z Internetu, co jest najbardziej poszukiwanym obszarem w dziedzinie danologii.

Ponieważ w serwisie LinkedIn pojawia się coraz więcej ofert pracy w dziedzinie danologii i przewiduje się, że do 2026 r. utworzy ona 11,5 miliona nowych miejsc pracy, osoby poszukujące pracy, które posiadają ten zestaw umiejętności, mają w tym zakresie szerokie możliwości. Celem tej książki jest dostarczenie odpowiedniej wiedzy z zakresu eksploracji sieci Web dla czytelników, którzy chcieliby zajmować się analizą danych.

Książka podzielona jest na trzy części: wprowadzenie i pojęcia, metodologie oraz zastosowania.

Rozdział 1: Eksploracja sieci Web – Wprowadzenie stanowi wprowadzenie do eksploracji sieci Web oraz omawia ewolucję i podstawowe pojęcia związane z eksploracją danych w sieci.

Rozdział 2: Taksonomia eksploracji sieci Web omawia taksonomię eksploracji sieci Web, eksplorację zawartości sieci Web, eksplorację struktury sieci Web, eksplorację korzystania z sieci Web oraz inne ważne pojęcia.

Rozdział 3: Główne zastosowania eksploracji sieci Web omawia podstawowe zastosowania eksploracji sieci Web, spersonalizowane aplikacje dla klientów, wyszukiwanie w sieci, śledzenie w sieci i eksplorację procesów.

Rozdział 4: Podstawy języka Python omawia różne zastosowania z metodologiami eksploracji i podstawami języka programowania Python. Obejmuje również podstawy Pythona, podstawowe znaczniki HTML i podstawy dotyczące bibliotek Pythona.

Rozdział 5: Ekstrakcja danych z sieci Web omawia zastosowania ekstrakcji danych z sieci, sposoby ekstrakcji danych, moduły języka Python, a także legalność wyodrębniania, ekstrakcji i wstępnego przetwarzania danych.

Rozdział 6: Eksploracja opinii omawia pojęcia dotyczące eksploracji opinii, przetwarzania i tokenizacji danych oraz wyodrębniania cech.

Rozdział 7: Eksploracja struktury sieci Web skupia się na koncepcji i rodzajach eksploracji struktury sieci Web. Omawia również eksplorację grafów, eksplorację sieci Deep Web, wyszukiwanie w sieci oraz hiperłącza.

Rozdział 8: Analiza sieci społecznych w języku Python stanowi wprowadzenie do analizy sieci społecznościowych, tworzenia sieci symetrycznych i asymetrycznych oraz łączności w sieci.

Rozdział 9: Eksploracja korzystania z sieci Web omawia źródła i rodzaje danych, kluczowe elementy wstępnego przetwarzania danych dotyczących korzystania z sieci, modelowanie danych oraz odkrywanie i analizowanie wzorców.

Web mining jest rozległym i bardzo aktywnym obszarem badań. Wszystkie koncepcje staraliśmy się wyjaśnić w taki sposób, aby książka nie stała się zbyt obszerna, a przy tym była prosta i zrozumiała. Chętnie zapoznamy się z opiniami ze strony naszych czytelników. Mamy nadzieję, że książka pomoże czytelnikom w poznaniu koncepcji związanych z eksploracją danych w sieci Web z użyciem języka Python.

Dr Ranjana Rajnish
Dr Meenakshi Srivastava

Rozdział 1

Eksploracja sieci Web – Wprowadzenie

Wstęp

Eksploracja sieci Web (Web mining) to proces polegający na odkrywaniu i wyodrębnianiu informacji ze stron internetowych przy użyciu różnych technik eksploracji danych. Te informacje mogą być wykorzystywane przez przedsiębiorstwa do efektywnego podejmowania decyzji. Ten rozdział stanowi wprowadzenie do sieci World Wide Web, wyjaśnia podstawy eksploracji danych i eksploracji sieci Web, a także omawia rodzaje informacji, jakie mogą być eksplorowane, oraz ich zastosowanie. Omawiany jest tu również sposób wykorzystania języka Python do eksploracji sieci Web. Rozdział przeznaczony jest dla początkujących osób, które są nowicjuszami w dziedzinie eksploracji sieci Web. Celem tego rozdziału jest dostarczenie obszernego wprowadzenia do tematu w celu ułatwienia zrozumienia kolejnych rozdziałów.

Struktura

W tym rozdziale omawiane są następujące tematy:

- Wprowadzenie do eksploracji sieci Web
- Sieć World Wide Web
- Internet i Web 2.0
- Wstęp do eksploracji, modelowania i analizy danych
- Ewolucja eksploracji sieci Web

- Podstawy eksploracji sieci Web
- Zastosowania eksploracji sieci Web
- Eksploracja sieci Web i język Python
- Podsumowanie
- Pytania i ćwiczenia

Cele

W ramach tego rozdziału poznasz podstawy eksploracji sieci Web, ewolucję sieci Web, podstawowe pojęcia związane z eksploracją danych w Internecie, a także różnice między eksploracją danych i eksploracją sieci Web. Dowiesz się również, dlaczego język Python jest pomocny przy eksplorowaniu sieci Web i jakie kroki są niezbędne do eksplorowania informacji.

Wprowadzenie do eksploracji sieci Web

Eksploracja sieci Web jest procesem polegającym na odkrywaniu i wyodrębnianiu informacji z Internetu za pomocą różnych technik eksplorowania danych. Te informacje mogą być wykorzystywane przez firmy do efektywnego podejmowania decyzji.

Dawniej dane przechowywane były w bazach danych i miały ustrukturyzowaną formę, dzięki czemu dowolne informacje można było pozyskać z użyciem zapytań do tych baz danych. Rozpowszechnianie informacji odbywało się wówczas przy wykorzystaniu raportów generowanych na podstawie informacji przechowywanych w bazie danych. Obecnie najpopularniejszą metodą szerzenia informacji jest sieć *World Wide Web* (WWW), w związku z czym przechowuje ona ogromne ilości informacji. Sieć Web 2.0 zmieniła sposób, w jaki postrzegamy dane. Z kolei sieć Web 3.0 jest siecią, której jedną z funkcji jest baza danych. Taka sieć daje nam możliwość eksplorowania Internetu jako wielkiej bazy danych, która jest wypełniona informacjami. Dzięki procesom odkrywania wiedzy (*Knowledge Discovery*, KD), z tej wielkiej bazy danych możemy wyodrębnić ważne dane zawierające różnorodne informacje, takie jak tekst, obrazy, wideo czy multimedia.

Dawno minęły już czasy, kiedy ludzie chodzili czytać do biblioteki. Gdy przychodzi nam do głowy jakieś zapytanie, zwykle wyszukujemy je w Internecie za pomocą dowolnej wyszukiwarki (takiej jak Google, Yahoo itd.). Przy ponad 560 milionach użytkowników Internetu, Indie zajmują drugie miejsce pod względem

wykorzystania Internetu, plasując się tuż za Chinami. To pozwala nam uświadomić sobie liczbę osób, które z różnych powodów korzystają z Internetu. Przy tak dużej ilości danych, jaka dostępna jest w Internecie, musimy je konwertować na istotne informacje, które mogą zostać użyte w jakimś sensownym zastosowaniu. Jednak samo pozyskanie danych to za mało, abyśmy mogli wykorzystać je w pełni. Potrzebna jest jeszcze jakaś metodologia, która pomoże nam wyodrębnić te dane z sieci i skonwertować je do postaci sensownych informacji.

Web mining jest procesem eksplorowania lub wyodrębniania istotnych informacji z Internetu. Dwie inne często używane definicje dla eksploracji sieci Web są następujące:

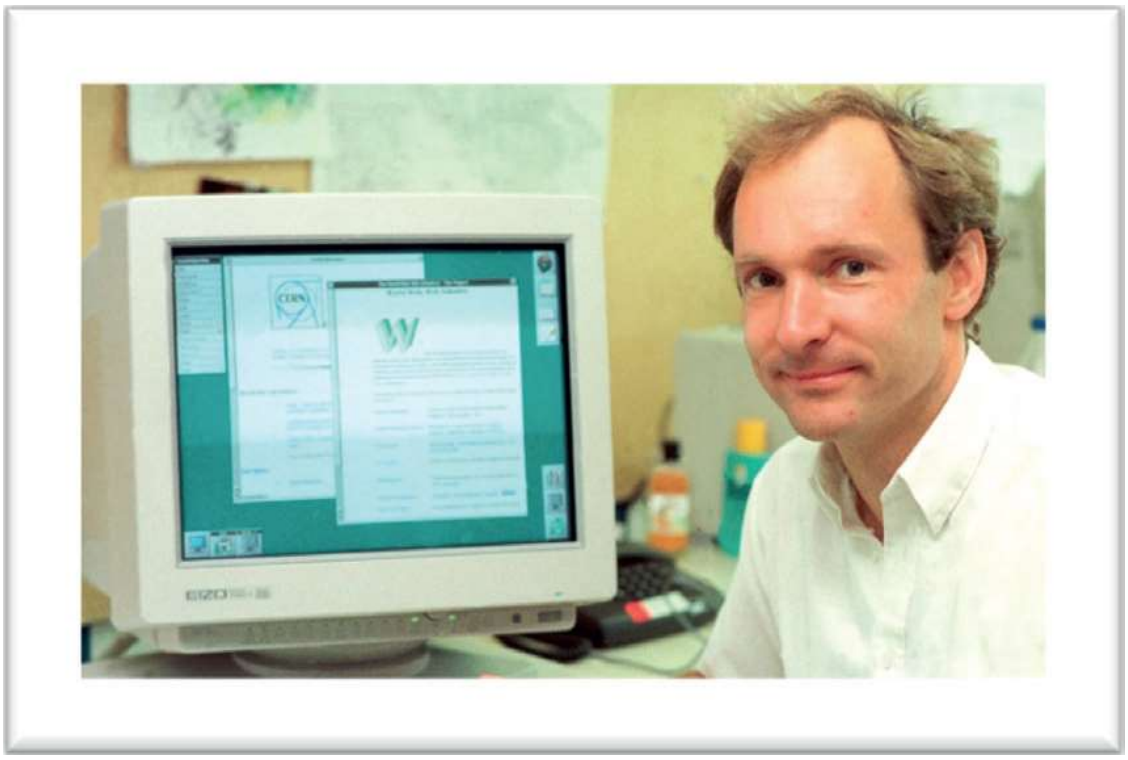
> *Web mining polega na wykorzystaniu technik eksploracji danych do automatycznego odkrywania i wyodrębniania informacji z dokumentów/usług w sieci Web* (*Etzioni, 1996, CACM 39*(*11*)).
>
> *Celem eksploracji sieci Web jest odkrywanie przydatnych informacji lub wiedzy ze struktury hiperłączy, zawartości stron i danych dotyczących wykorzystania* (*Bing LIU 2007, Web Data Mining, Springer*).

Sieć World Wide Web

Sieć World Wide Web, powszechnie nazywana siecią WWW lub internetem, miała swój skromny początek w CERN, międzynarodowej organizacji naukowej w Genewie w Szwajcarii, gdzie w 1989 r. stworzył ją brytyjski naukowiec Tim Berners-Lee. Jego zdaniem udostępnianie informacji było trudne, ponieważ trzeba było w tym celu logować się do różnych komputerów. Myślał więc nad rozwiązaniem tego problemu i w marcu 1989 r. przedstawił swoją wstępną propozycję o nazwie *Information Management: A proposal* (Zarządzanie informacją: propozycja).

W 1990 r. sformalizował tę propozycję i razem z belgijskim inżynierem systemowym Robertem Cailliauem przedstawił jej kolejną wersję. W tej propozycji nakreślił pojęcia związane z siecią Web i nazwał ją „projektem hipertekstowym" o nazwie „WorldWideWeb". Wspólnie zaproponowali, że sieć ta będzie składać się z „dokumentów hipertekstowych" możliwych do wyświetlania w przeglądarkach. Tim Berners-Lee opracował pierwszą wersję tej sieci z działającym serwerem sieci Web i przeglądarką, demonstrując w ten sposób idee zaprezentowane w ich propozycji. Adresem pierwszej witryny internetowej było *info.cern.ch*, a sama witryna była hostowana na komputerze NeXT w CERN.

Witryna zawierała informacje o projekcie WWW, wraz ze wszystkimi szczegółami na jego temat. Adresem pierwszej strony internetowej było *http://info.cern.ch/hypertext/WWW/TheProject.html*. Aby mieć pewność, że komputer używany jako „serwer sieci Web" nie zostanie przypadkowo wyłączony, wyraźnie napisano czerwonym kolorem: „This machine is a server. DO NOT POWER IT DOWN" (Ten komputer jest serwerem. NIE WYŁĄCZAJ GO).

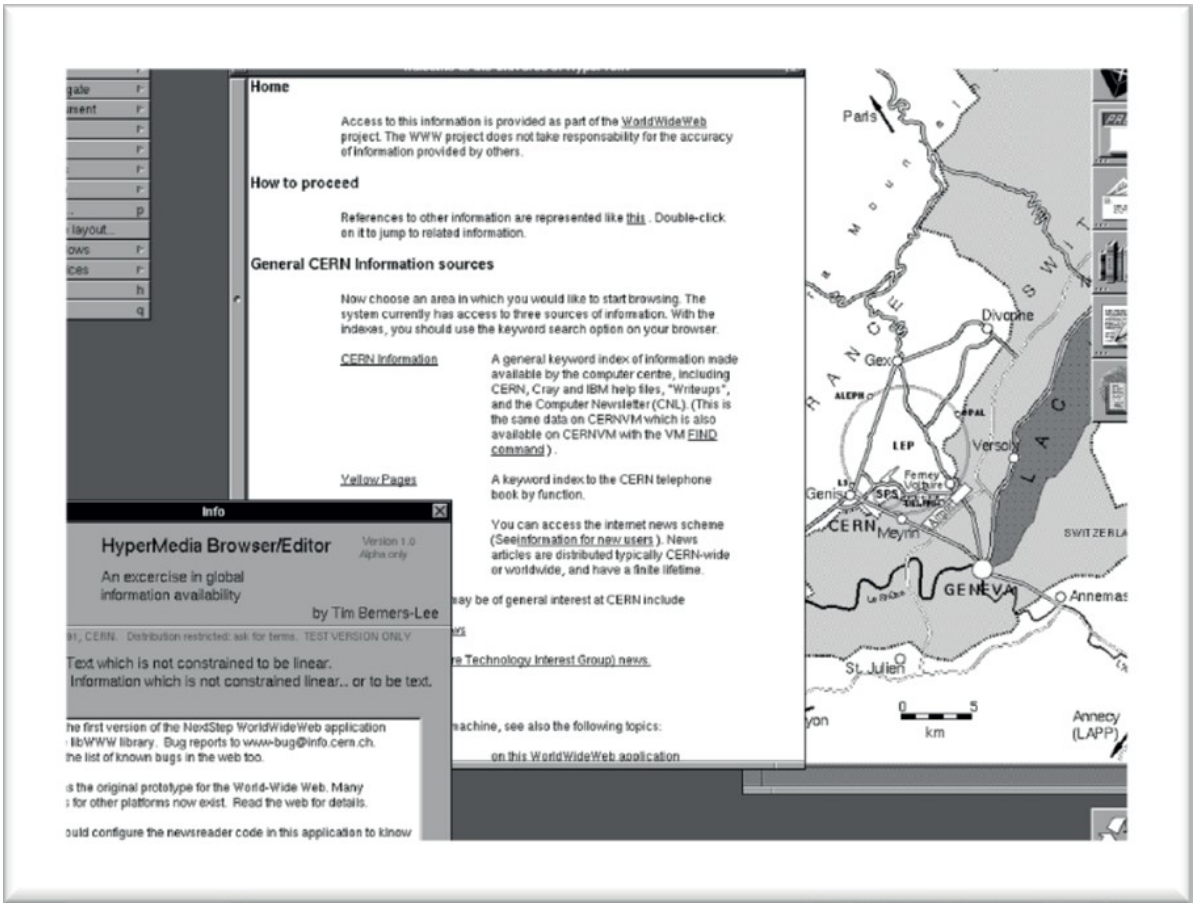


Rysunek 1.1: *Tim Berners-Lee w CERN (Zdjęcie: CERN)*
Źródło: https://www.britannica.com/topic/World-Wide-Web

Pierwotnie sieć Web została wymyślona i opracowana w celu automatycznej wymiany informacji między naukowcami pracującymi na uniwersytetach i w instytutach na całym świecie. Rysunek 1.2 jest zrzutem ekranu pokazującym przeglądarkę NeXT World Wide Web.

Witryna zapewniała łatwy dostęp do istniejących informacji przydatnych naukowcom CERN. Oferowała funkcjonalność wyszukiwania przy użyciu prostych słów kluczowych, gdyż w tamtym czasie nie istniała jeszcze żadna wyszukiwarka. Ten projekt miał na początku ograniczoną funkcjonalność i tylko kilku użytkowników miało dostęp do platformy komputerowej NeXT (użytej do stworzenia serwera). W marcu 1991 r. projekt został udostępniony wszystkim

współpracownikom korzystającym z komputerów w CERN. W sierpniu 1991 r. Berners-Lee ogłosił wersję WWW tej sieci na internetowych grupach dyskusyjnych, dzięki czemu projekt rozprzestrzenił się po całym świecie.



Rysunek 1.2: *Zrzut ekranu ukazujący przeglądarkę NeXT World Wide Web, którą stworzył Tim Berners-Lee*

Zdjęcie: CERN

Niedługo po tym Komisja Europejska połączyła siły z CERN, po czym CERN udostępnił bezpłatnie kod źródłowy serwera World Wide Web. Pod koniec 1993 r. działało już ponad 500 znanych serwerów sieci Web, a sieć WWW stanowiła 1% całego ruchu internetowego (za pozostały ruch była odpowiedzialna poczta internetowa, dostęp zdalny i przesyłanie plików).

Tim zdefiniował trzy podstawowe elementy składowe Internetu – HTML, URI i HTTP – które stanowią podstawę dzisiejszej sieci Web.

Hyper Text Markup Language (HTML) używany jest jako język znaczników (formatowania).

Uniform Resource Identifier (URI), znany również jako Uniform Resource Locator (URL), używany jest jako unikatowy adres do lokalizowania każdego zasobu w sieci Web.

Hypertext Transfer Protocol (**HTTP**) jest protokołem, który pomaga pozyskiwać podlinkowane zasoby z Internetu.

W 1995 r. popularność sieci Web znacząco wzrosła, gdy korporacja Microsoft, gigant w obszarze oprogramowania, zaczęła wspierać aplikacje internetowe na komputerach osobistych i opracowała swoją własną przeglądarkę, Internet Explorer (IE), która była początkowo oparta na przeglądarce Mosaic. W 1996 r. Microsoft zintegrował przeglądarkę IE z systemem operacyjnym Windows, przez co IE stała się najpopularniejszą przeglądarką internetową.

Ewolucja sieci World Wide Web

Od czasu jej powstania sieć World Wide Web przeszła ogromną zmianę. Każdy etap jej ewolucji wniósł do niej sporą wartość i można go kategorycznie wyróżnić odrębnymi pojęciami. W tym podrozdziale omówimy krótko ewolucję sieci Web.

Pod koniec 1994 r. w użyciu było około 10 000 serwerów, z czego 2000 wykorzystywano komercyjnie. Sieć Web była wtedy używana przez ponad 10 milionów użytkowników, a ruch internetowy znacząco się zwiększył. Technologia była stale badana, aby móc zaspokajać inne potrzeby, takie jak narzędzia związane z bezpieczeństwem, handel elektroniczny i aplikacje.

Pierwotnie wersja podstawowa sieci, czyli **Web 1.0** (1989), zaprojektowana była pod kątem publikowania informacji, które mogliby czytać wszyscy. Ta era charakteryzowała się hostingiem stron informacyjnych zawierających informacje korporacyjne, takie jak informacje organizacyjne, broszury itd., pomagając w rozwoju biznesu. Można więc powiedzieć, że był to zbiór olbrzymiej liczby dokumentów, które można było czytać w obrębie całej sieci World Wide Web. Głównym celem w tej erze było stworzenie wspólnego miejsca, w którym można byłoby dzielić się informacjami. W tej erze sieć Web istniała jako sieć tylko do odczytu i składała się ze statycznych stron HTML.

> ***Uwaga:*** Sieć Web 1.0 została zaprojektowana w celu wymiany informacji i pozwalała jedynie na publikowanie informacji na stronach. Użytkownik mógł jedynie odczytywać te informacje.

W roku 2004 rozwinęła się sieć **Web 2.0**, znana jako sieć zorientowana na ludzi, sieć partycypacyjna lub sieć społecznościowa. Ponieważ sieć stała się dwukierunkowa, pozwalając na wykonywanie operacji odczytu i zapisu, zaczęła być używana jako platforma do współpracy i stała się interaktywna. Zastosowane w tej wersji technologie upraszczały wymianę informacji, zapewniały interoperacyjność, wspierały projektowanie zorientowane na użytkownika i ułatwiały współpracę. W tym czasie powstały takie usługi i serwisy, jak strony wiki, blogi, YouTube, Facebook, LinkedIn czy Wikipedia. Tak więc ta era charakteryzowała się zarówno odczytem, jak i zapisem. Łączyła ona nie tylko dokumenty, ale także użytkowników.

Uwaga: Sieć Web 2.0 to sieć zorientowana na ludzi, sieć partycypacyjna lub sieć społecznościowa, w której użytkownicy mogą zarówno odczytywać, jak i zapisywać informacje.

Web 3.0, czyli sieć Web trzeciej generacji, została stworzona jako sieć, która pomaga w bardziej efektywnym odkrywaniu, automatyzacji i integracji, ponieważ łączy ludzką i sztuczną inteligencję w celu dostarczania bardziej istotnych informacji. Sieć Web 3.0 kładzie nacisk na analizę, przetwarzanie i generowanie nowych idei w oparciu o informacje dostępne w obrębie sieci. Sieć Web 3.0 ukuła nową koncepcję polegającą na przekształceniu sieci Web w bazę danych, czyniąc ją w ten sposób bardziej przydatną dla takich rozwiązań, jak sztuczna inteligencja, grafika 3D, łączność, wszechobecność i sieć semantyczna. Jest ona również znana jako semantyczna sieć Web i również została stworzona przez Tima Bernersa-Lee, twórcę sieci Web 1.0. Sieć semantyczna kładzie nacisk na uczynienie sieci czytelną dla maszyn i zdolną do odpowiadania na złożone zapytania zadawane przez ludzi na podstawie ich znaczenia. Według W3C, „semantyczna sieć Web zapewnia wspólną strukturę, która umożliwia udostępnianie i ponowne wykorzystywanie danych między aplikacjami, przedsiębiorstwami i społecznościami".

Uwaga: Sieć Web 3.0 charakteryzują takie rozwiązania, jak sztuczna inteligencja, grafika 3D, łączność, wszechobecność i sieć semantyczna.

Sieć **Web 4.0** jest bardziej rewolucyjna i oparta na komunikacji bezprzewodowej (mobilnej lub komputerowej), w ramach której ludzie mogą łączyć się

z obiektami. Przykładowo samochody wyposażone w moduł GPS pomagają kierowcy wybrać najkrótszą możliwą trasę. Ta generacja określana jest mianem „inteligentnej sieci Web" i będzie widoczna w latach 2020 – 2030. W tej generacji komputery odgrywają różne role, od asystentów osobistych po wirtualne rzeczywistości. Sieć będzie miała inteligencję podobną do ludzkiej, a w jej obrębie będzie zachodzić wysoce inteligentna interakcja między ludźmi i maszynami. Do sieci będą mogły zostać podłączone wszystkie urządzenia gospodarstwa domowego, a nawet będą prowadzone prace nad implantami mózgowymi.

> ***Uwaga:*** Sieć Web 4.0 postrzegana jest jako „mobilna sieć Web" lub „inteligentna sieć Web".

Sieć **Web 5.0** jest futurystyczną siecią Web, nad którą prowadzi się obszerne badania. Przewiduje się, że będzie to „telepatyczna i emocjonalna sieć Web", która powinna pojawić się do 2030 roku.

Internet i Web 2.0

Internet oraz pojawienie się sieci Web 2.0, znanej jako sieć zorientowana na ludzi, sieć partycypacyjna lub sieć społecznościowa, zapewniło nowe sposoby wykorzystywania informacji w Internecie z korzyścią dla społeczeństwa. Był to czas, kiedy nastąpiła zasadnicza zmiana w sposobie korzystania z Internetu. Wcześniej Internet używany był jako narzędzie. Wraz z nadejściem sieci Web 2.0 Internet stał się częścią naszego życia, przekształcając się ze statycznej sieci Web w sieć społecznościową. W tej erze nie tylko zwiększyliśmy ilość wykorzystywanych danych, ale również wydłużyliśmy czas korzystania z Internetu. Strony internetowe stały się bardziej interaktywne, a nowe technologie umożliwiły stronom internetowym oddziaływanie z przeglądarką internetową bez ludzkiej interwencji.

Odkąd zaczęliśmy korzystać z różnych inteligentnych mediów, takich jak smartfony, tablety, laptopy i odtwarzacze MP3, różnych narzędzi, takich jak wyszukiwarki internetowe (np. Google i Yahoo), narzędzi do udostępniania zdjęć i filmów (np. YouTube i Instagram) czy mediów społecznościowych (np. Facebook i WhatsApp), Internet stał się integralną częścią naszego życia. Za pośrednictwem różnych platform generowane są olbrzymie ilości danych w postaci tekstu, obrazów czy filmów. Skutkiem tego jest przeciążenie informacjami.

Z tego względu ważne stało się wyodrębnianie sensownych i istotnych informacji z dużych ilości danych dostępnych w Internecie. Tak narodziły się technologie i rozwiązania do wyszukiwania informacji, takie jak web mining.

Eksplorowanie, modelowanie i analizowanie danych

Rozwój internetowych aplikacji biznesowych we wszystkich branżach oraz automatyczne generowanie danych za pośrednictwem różnych źródeł w Internecie doprowadziło do powstania olbrzymich repozytoriów danych. Sieci handlowe, takie jak Walmart i Big Bazar, mają tysiące sklepów z milionami dziennych transakcji. W celu zarządzania sposobem przechowywania tak wielkich ilości danych wprowadzono wiele innowacji technicznych. Z uwagi na konieczność zarządzania tymi danymi nastąpił szybki rozwój technologii zarządzania bazami danych, ale metodologie używane do pozyskiwania i analizy tych danych były trywialne. Dopiero kiedy firmy zaczęły zdawać sobie sprawę, że w tym ogromie surowych danych jest wiele od odkrycia, informatycy zaczęli pracować nad sposobami eksploracji tych ukrytych informacji. Ogromna ilość danych zawierała wiele ukrytych faktów lub wzorców, które można było zbadać w celu umożliwienia stworzenia lepszych systemów wspomagania decyzji pozwalających na podejmowanie bardziej skutecznych decyzji. Dane skrywały w sobie szeroką wiedzę na temat kilku aspektów związanych z biznesem, które można było wykorzystać do skutecznego i efektywnego podejmowania decyzji. To wydobywanie wiedzy z baz danych lub zbiorów danych znane jest jako eksploracja danych (Data mining) lub odkrywanie wiedzy w bazach danych (Knowledge Discovery in Databases, KDD).

Czym jest eksploracja danych?

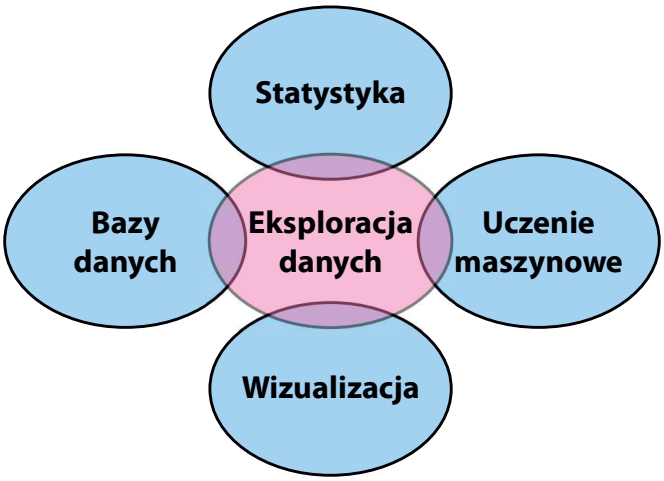
Według Gartnera, „Eksploracja danych jest procesem odkrywania nowych istotnych korelacji, wzorców i trendów poprzez przeczesywanie dużych ilości danych przechowywanych w repozytoriach z użyciem technologii rozpoznawania wzorców oraz technik statystycznych i matematycznych".

Rozumiejąc potencjał eksploracji danych, opracowano wiele technologii, które pomagają analizować te olbrzymie ilości danych, często określanych pojęciem Big Data. Na rynku uwaga przeniosła się z produktów na klientów, a trend zmienił się w kierunku dostarczania spersonalizowanych transakcji. Technologia

używana do przechwytywania danych również przeszła z ręcznej na zautomatyzowaną z wykorzystaniem kodów kreskowych, terminali Point of Sale i tak dalej. Technologie zarządzania bazą danych były początkowo używane do wydajnego przechowywania, wyszukiwania i manipulowania danymi, ale w związku z nowym wymogiem eksploracji danych opracowano wiele algorytmów do eksplorowania tych informacji. W tym czasie zaczęło ewoluować także uczenie się maszyn, a dzięki połączeniu technik eksploracji danych i algorytmów uczenia się maszyn nastąpiła rewolucja w dziedzinie eksploracji.

> ***Uwaga:*** Pojęcie Big Data odnosi się do olbrzymich ilości danych, charakteryzowanych przez ilość, szybkość i wiarygodność. Można je analizować obliczeniowo w celu znalezienia ukrytych wzorców, trendów lub powiązań.

Eksploracja danych wykorzystuje koncepcje znane z technologii baz danych, statystyki, uczenia się maszyn, wizualizacji i grupowania. Przedstawia to rysunek 1.3:



Rysunek 1.3: *Koncepcje wykorzystywane w eksploracji danych*

Czym jest, a czym nie jest eksploracja danych? Wielu użytkowników ma wątpliwości co do tego, czym eksploracja danych różni się od zwykłego wyszukiwania w bazie danych. Spójrzmy na kilka przykładów, które pomogą nam to zrozumieć:

Wyszukiwanie numeru telefonu w książce telefonicznej *nie jest eksploracją danych.*

Wyszukiwanie studentów, którzy uzyskali oceny na poziomie powyżej 75% *nie jest eksploracją danych.*