POZYSKIWANIE I PRZECHOWYWANIE DANYCH 7

Przedstawione dotychczas przykłady sieci powstały z wykorzystaniem ogólnodostępnych źródeł danych. Dane takie trzeba jednak najpierw odszukać, a następnie pobrać i przechować. W tym rozdziale omawiamy techniki pozyskiwania i magazynowania informacji, które uznaliśmy za wyjątkowo użyteczne. Niektóre mogą wydawać się anachroniczne (analiza WWW, bazy relacyjne), lecz zapewniamy, że są one wciąż bardzo przydatne na co dzień.

7.1. Analiza stron WWW

Strony WWW jako źródło danych

Wszyscy codziennie korzystamy z zasobów internetu, co wynika z naszej świadomości, że informacje, które znajdujemy w rozlicznych witrynach internetowych, mogą być dla nas przydatne. Często też mamy do czynienia z stronami zawierającymi znaczną, a bywa, że wręcz przytłaczającą ilość informacji. Może pojawić się wówczas myśl, że warto było by te informacje w jakiś sposób przetworzyć, gdyż z jednej strony jest ich zbyt dużo, aby mógł zrobić to przeciętny użytkownik, z drugiej zaś z ich obfitości wynika ich wartość. W końcu w wyniku analizy informacji pochodzących z różnych źródeł może powstać nowa jakość, można też uzyskać wiedzę wcześniej niedostępną a potrzebną.

Najprostszym przykładem mogą być różnego rodzaju serwisy integrujące informacje: wyszukujące najtańsze oferty, informujące o wydarzeniach kulturalnych czy imprezach rozrywkowych. Oczywiście część, jeśli nie większość z nich, prezentuje dane pozyskane bezpośrednio od zainteresowanych dostawców treści, często za pośrednictwem w tym celu zaprojektowanych interfejsów programistycznych. Ich wykorzystanie

stanowi istotne zagadnienie, będziemy o tym jeszcze pisać w dalszej części rozdziału. Tymczasem zajmiemy się pobieraniem danych ze stron, które nie są specjalnie do tego celu przeznaczone, innymi słowy – są zaprojektowane tak, aby użytkownik mógł z nich wygodnie korzystać przy użyciu przeglądarki.

Należy tutaj zwrócić uwagę na dwa aspekty takiego zdobywania danych: techniczny i prawny. Z czysto technicznego punktu widzenia należy napisać program, który będzie naśladował użytkownika, czy raczej użytkownika i przeglądarkę, wczytywał dane, korzystając z protokołu HTTP, obrabiał je, a następnie zapisywał w odpowiedni sposób. Jest to zadanie trudne, gdyż jak już zauważyliśmy, typowe strony WWW nie są przeznaczone do takiego wykorzystania, mogą zawierać wiele zbędnych, można powiedzieć zaciemniających dane elementów, mogą też być wyposażone w mechanizmy przeciwdziałające nadmiernemu obciążaniu serwera, czy wręcz automatycznemu przeglądaniu. Z drugiej strony kod HTML jest przeznaczony do interpretowania przez przeglądarkę, jest więc również możliwe stworzenie takiego interpretera, który wydobędzie poszukiwane dane, co więcej oprogramowanie takie istnieje w postaci zarówno gotowych programów, jak i bibliotek – jedną z nich zaprezentujemy w skrócie w tym podrozdziale.

Zarówno z prawnego, jak również moralnego punktu widzenia musimy być przede wszystkim świadomi poszanowania praw autorskich i majątkowych do pozyskiwanych danych. Często spotykane podejście, polegające na traktowaniu jako własność publiczną wszystkich informacji znajdujących się w sieci i nie obwarowanych restrykcjami np. w postaci płatnego dostępu, jest z gruntu błędne. Wiele z nich jest bowiem prezentowane nam przez ich właściciela w ściśle określonym celu, np. umieszczenie w ogólnie dostępnym serwisie streszczenia sztuki teatralnej służy poinformowaniu potencjalnych widzów o jej treści, nie jest jednak jednoznaczne z powszechnym zezwoleniem na jego powtórną publikację w innych serwisach czy wydawnictwach. Nie wgłębiając się w szczegóły prawne, mamy tu sytuację podobną jak w przypadku zakupu książki – zyskujemy wtedy prawo własności konkretnego egzemplarza, który możemy przeczytać, a tak zdobyte informacje wykorzystać, np. aby podnieść swoje umiejętności. Nie zdobywany jednak prawa do przedruku – te prawo pozostaje zazwyczaj przy autorze lub wydawnictwie, jeśli autor mu swoje prawa przekazał.

Powstaje zatem pytanie, czy w tej sytuacji możliwe jest wykorzystanie jakichkolwiek danych. Odpowiedź twierdzącą można z całą pewnością dać w przypadku danych pozostających w domenie publicznej – znajdziemy wtedy informację, że są one dostępne i można je ponownie publikować. Danych takich w sieci jest coraz więcej i często są one bardzo wartościowe dzięki pracy wielu osób zaangażowanych w ich tworzenie. Przykładem mogą być np. WikiŹródła czy OpenStreetMap. W innych przypadkach należy poszukiwać klauzuli o możliwości wykorzystania danych, można też próbować skontaktować się z właścicielem serwisu w celu uzyskania pozwolenia. W szczególności instytucje państwowe udostępniają wiele danych, które mogą być wykorzystywane po spełnieniu pewnych, często bardzo prostych warunków. Wreszcie, zgodnie z prawem obowiązującym w Polsce prawa autorskie przedawniają się po upływie 70 lat od chwili śmierci ich autora co otwiera dostęp do wielu historycznych utworów i dokumentów.