

Zbiorowa inteligencja rekomenduje

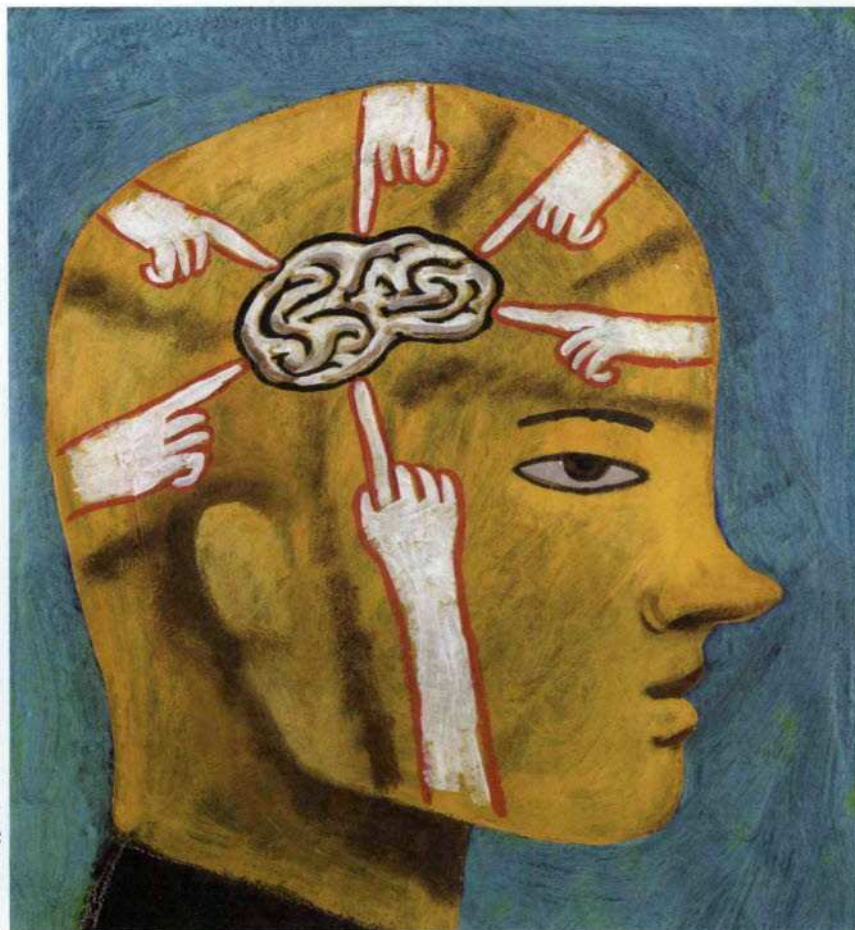
tekst: Marcin Majda

Wypożyczalnia NetFlix ogłosiła w 2006 r. konkurs, w którym zaoferowała milion dolarów dla drużyny, która poprawi ich algorytm rekomendacji o 10 proc. Dlaczego tak duża nagroda i taki konkurs? Rekomendacje dla NetFlix są kluczowym elementem, gdyż odpowiadają za 60 proc. sprzedaży!

Miliony ludzi, tworząc własne strony www, decydują się na umieszczenie odnośników do innych stron, które uważają za ważne i interesujące. Im więcej osób zamieści na swojej stronie link do danej strony, tym wyżej zostanie ona oceniona. Całkowicie nowe podejście, wykorzystujące informacje wygenerowane przez miliony ludzi, pozwoliło firmie Google zdominować rynek wyszukiwarek, a jej autorzy znajdują się na liście najbogatszych ludzi świata.

Kolejnym przykładem wykorzystania wiedzy wielu jednostek jest Wikipedia, która jest internetową encyklopedią tworzoną i generowaną przez użytkowników. Sama strona niewiele robi, pozwala dodawać, edytować wpisy oraz wyszukiwać interesujące nas informacje. Jednak dzięki zbiorowej współpracy wielu ludzi udało się stworzyć tak dużą encyklopedię, której żaden zespół redakcyjny na świecie nie jest w stanie przygotować.

Zbiorowa inteligencja może być również wykorzystywana do szybkiego wykrywania informacji na temat zagrożeń, jakimi mogą być powodzie, burze, pożary.



Przykładem takiego rozwiązania jest projekt WeKnowIt (www.weknowit.eu) stworzony przez konsorcjum sześciu ośrodków naukowych z całej Europy oraz firm: Vodafone (Grecja), Telefonica (Hiszpania), Yahoo (USA) oraz krakowski Software Mind. Fundamentem projektu jest wykorzystanie wiedzy poszczególnych jednostek. Każdy obywatel na wiele sposobów, m.in. za pomocą SMS-a, MMS-a, nagrania wideo może informować o istniejących zagrożeniach. Zbieranie oraz wielowymiarowa analiza danych gromadzonych przez społeczeństwo może dostarczyć w krótkim czasie ogromnie ważnych informacji dla ośrodków koordynujących kataklizm. Najlepszym przykładem może być informacja o tym, że wszystkie drogi dojazdowe do szpitala są zablokowane, co w sytuacji kryzysowej może uratować życie wielu ludziom.

Jednak najpopularniejszym przykładem zbiorowej inteligencji omawianej w kontekście analizy zachowań mas ludzi są systemy rekomendacyjne oraz personalizacja. Bardzo często spotykamy się z nimi w serwisach internetowych. Systemy te podpowiadają nam automatycznie, bez udziału pracowników serwisu, co może się nam spodobać, starając się wpasować w nasz gust i spersonalizować zawartość strony specjalnie dla nas w czasie, kiedy ją oglądamy, oceniając nasze preferencje za pomocą algorytmów zbiorowej inteligencji.

Mechanizmy stosowane w praktyce

W sklepach internetowych najczęstszy przykład silnika rekomendacji jest oparty na śledzeniu pozycji, które kupowali użytkownicy. W takich serwisach jest osobna sekcja z rekomendacjami dla nowego użytkownika zawierająca wiedzę o preferencjach innych osób o podobnych do niego upodobaniach. Sekcje te są zatytułowane: „Osoby, które kupowały tę pozycję, często

Rys. 1. Wygląd rekomendacji w amazon.com „Ludzie, którzy kupili tę pozycję, kupili również...”



kupowały również...”. Jest to najprostszy mechanizm (tzw. Collaborative Filtering) i najczęściej wykorzystywany. Największy na świecie sklep internetowy amazon.com na tym nie poprzestaje. Dlaczego? Jak twierdzi Greg Linden, autor systemu rekomendacyjnego amazon.com, kiedy odchodził z firmy w 2002 roku ponad 20 proc. sprzedaży pochodziło z personalizacji i rekomendacji. Obecnie (jak podaje na swoim blogu) 35 proc. sprzedanych pozycji na amazon.com pochodzi z rekomendacji.

Jakie dodatkowe mechanizmy stosuje amazon.com? Oczywiście jest to pilnie strzeżona tajemnica firmy, jednak wnikliwa analiza serwisu ujawnia więcej detali. Na przykład, na każdej stronie produktu widzimy dodatkowe sekcje. Pierwszą z nich jest „Użytkownicy, którzy oglądali tę pozycję, ostatecznie kupili...”. Mamy

tutaj do czynienia ze śledzeniem, jakie poszczególne pozycje ogląda użytkownik. Schemat poruszania się użytkownika po serwisie kończy się często zakupem pewnej pozycji. Dostajemy do zanalizowania zachowania użytkowników, na podstawie których staramy się wydobyć informację, co może spodobać się naszemu klientowi. Użytkownik oglądał wiele pozycji, czytał ich opisy, opinie o nich, recenzje, aż w końcu na coś się zdecydował. Analizując tego typu zachowania, możemy wyciągać bardzo cenne informacje. Cenne o tyle, że dobrze dopasowane do użytkownika zwiększają szansę, że kupi on proponowany produkt, co jest istotne, bo ma on nikłe szanse znalezienia go w milionie innych produktów w serwisie. Wskazując nastolatce preparat do protezy, raczej nie osiągniemy sukcesu, w odróżnieniu od sytuacji, kiedy pokażemy jej

» produkt kupiony przez innego podobnego użytkownika (inną nastolatkę), np. najnowszą płytę popularnego wykonawcy. Inną sekcją pojawiającą się w księgarni amazon.com na stronie książki na temat „Collective Intelligence” jest „Poszukujesz pozycji Collective Intelligence?”. Tutaj mamy do czynienia z analizą języka naturalnego i z tego, co wiadomo, amazon.com stworzył własny system SIPs (Statistically Improbable Phrases). Polega on na znajdowaniu w danej książce wyrażen, które są najmniej prawdopodobne do znalezienia w innych książkach. W ten sposób budowane są grupy podobnych powiązanych pozycji.

Zastosowanie rekomendacji i personalizacji

Analiza ruchu użytkowników, porównywanie pozycji są wykorzystywane nie tylko w sklepach internetowych. Rekomendacje i personalizacja mają zastosowanie wszędzie tam, gdzie chcemy zwiększyć sprzedaż, czas przebywania użytkownika na stronie, jego zadowolenie oraz dopasować ofertę indywidualnie do potrzeb konkretnej osoby. Przed każdym wdrażanym systemem powinien być postawiony jasny cel biznesowy, który należy realizować. Oprócz samych systemów z rekomendacjami można wykorzystywać różnego rodzaju filtry modyfikujące wyniki pod konkretne potrzeby danej firmy. Możemy wyłączyć z rekomendowanych pozycji filmy erotyczne. Możemy rekomendować pozycje najbliższe użytkownikowi, ale wyłącznie z określonej kategorii, z tej, na której mamy największą marżę.

Drugim przykładem systemu, w którym personalizacja ma ogromne znaczenie, jest Netflix, największa wypożyczalnia filmów DVD w USA. Wypożycza

filmy i podczas tego procesu zbiera o nich recenzje od swoich klientów. Użytkownik w skali od 1 do 5 może ocenić każdą pozycję. Na podstawie danych o preferencjach wszystkich użytkowników system informatyczny jest w stanie zarekomendować internaucie film, o którym nigdy nie słyszał, a który mu się najprawdopodobniej spodoba. Wypożyczalnia Netflix ogłosiła w 2006 r. konkurs, w którym zaoferowała milion dolarów dla drużyny, która poprawi ich algorytm rekomendacji o 10 proc. Do tej pory, prawie po trzech latach, nie udało się tego dokonać żadnemu zespołowi, choć obecnie najlepsza drużyna w rankingu poprawia algorytm o 9,65 proc. W styczniu 2007 roku polski zespół z krakowskiej firmy Personal TV zajmującej się rekomendacjami osiągnął wynik lepszy niż osiąga Netflix o 4,36 proc., zajmując 25 miejsce na ponad 15 000 drużyn. Dlaczego zdecydowano się na tak dużą nagrodę i na taki konkurs? Rekomendacje dla Netflix są kluczowym elementem, gdyż odpowiadają za 60 proc. sprzedaży. Firma sama nieraz podejmowała próby ulepszenia algorytmu i wie, że jest to zadanie bardzo trudne, de facto przeznaczone dla zespołu naukowców. Reed Hastings, założyciel i prezes wypożyczalni stwierdził, że wprowadzenie algorytmu poprawiającego skuteczność rekomendacji o 10 proc. przyniesie tyle zysku, że jest na pewno warte dla firmy milion dolarów. Wynika to z tego, że lepsze rekomendacje przekładają się wprost na większą ilość wypożyczeń, gdyż klienci chętniej wypożyczają to, co rzeczywiście pasuje do ich gustów.

Opisywane powyżej modele opierają się na prostych algorytmach opartych na szukaniu podobnych pozycji (filmów czy książek) w dużej grupie danych. Amazon.com musi tak robić ze względu na skalę firmy. Posiada w swojej ofercie kilka milionów produktów i kilkadziesiąt milionów klientów.

Porównywanie użytkowników jako źródło rekomendacji

Przy mniejszych zastosowaniach (lub w systemach, które nie potrzebują odpowiedzi w czasie rzeczywistym) można wykorzystywać wyszukiwanie podobieństw między użytkownikami. W takich sytuacjach można zacząć dodatkowo wykorzystywać bardziej szczegółowe informacje o użytkowniku, takie jak: wiek, płeć, miejsce zamieszkania, wykształcenie itp. System rekomendacyjny nie musi wiedzieć, kim dana osoba dokładnie jest, potrzebny jej jedynie ustalony unikalny identyfikator oraz zbiór atrybutów związanych z nim lub jego dotychczasowymi preferencjami. Posiadając tego typu dane, można zacząć porównywać użytkowników i budować ciekawe wnioski, na przykład: „mężczyźni z wykształceniem wyższym w wieku 30-40 lat w 60 proc. kupują pozycje...”. Na tej podstawie można personalizować treści idealnie pod daną osobę, trafiając bardzo dobrze w jej gust. Zamiast porównywać przedmioty (item-based), przechodzimy do porównywania użytkowników (user-based) i w ten sposób generujemy rekomendacje. Stosowane są też modele klastrowe, w których łączymy użytkowników w grupy o podobnych preferencjach i przygotowujemy rekomendacje specjalnie dla nich. Bardzo często metody rekomendacji są łączone celem dostarczenia jak najlepszego wyniku i zależne są od konkretnego systemu, ilości posiadanych danych oraz stawianych celów biznesowych.

Warto również zwrócić uwagę na inne przykłady zastosowań zbiorowej inteligencji czy analizy zachowań użytkowników systemów informatycznych. Pierwsze z nich to serwisy wideo, takie jak YouTube.com, gdzie przy każdym oglądanym materiale mamy sekcję „Podobne filmy wideo”. Sekcja ta stara się podpowiadać, co jeszcze o podobnej tematyce do aktualnie prezentowanej pozycji możemy obej-

rzeń. Z jednej strony użytkownik dostaje więcej informacji na interesujący go temat (i co ważne, nie musi jej szukać w gąszczu innych pozycji), a z drugiej dzięki temu dłużej pozostanie na serwisie.

Kolejny przykład to systemy randkowe. Na podstawie preferencji i zachowań danego użytkownika serwis może polecać osoby, którymi byłby potencjalnie zainteresowany. Trzeci przykład to reklama behawioralna. Analizując w serwisach internetowych zachowania ludzi, sprawdzając, po jakich stronach chodzą, w które reklamy klikają, możemy budować ich profil, a w efekcie dostosowywać reklamy do ich preferencji.

Przewidywanie i prognozowanie

Istnieją na świecie serwisy, które analizują oferty mieszkań wystawionych na sprzedaż. Biorą pod uwagę wiele atrybutów, takich jak cena, wielkość mieszkania, liczba pokoi, lokalizacja, wiek bloku, piętro mieszkania. Na podstawie tych informacji są w stanie prognozować ceny mieszkań niewystawionych jeszcze na sprzedaż.

Aktywizacja tzw. długiego ogona, czyli prezentacja najmniej popularnych pozycji, jest problemem wielu serwisów internetowych.

W serwisach wideo, w sklepach internetowych często mamy do czynienia z listą TOP 10 – najpopularniejszych, najlepszych pozycji. Jak mówi zasada Pareto, 80 proc. użytkowników ogląda 20 proc. produktów. To źle, ponieważ pozostała część pozycji też jest ciekawa, ale przeciętny użytkownik nie przejrzy kilkuset produktów, nie ma na to siły. Wystarczy jednak zaproponować mu 10 najbardziej do niego pasujących, i jeśli rekomendacja będzie trafna, kilka z nich kupi. Aktywizacja długiego ogona jest zaprzeczeniem zasady Pareto. Systemy rekomendacyjne, personalizacyjne, porównując zawartości produktów, analizując zachowania ludzi dążą do przedstawienia użytkownikowi pozycji bardzo mało popularnych, ale razem takich, które mu się spodobają i z których będzie zadowolony. W ten sposób pełna oferta docierająca do zainteresowanych osób.

Personalizacja, rekomendacje, wielowymiarowa analiza danych opierają się na elementach sztucznej inteligencji. Wykorzystują takie rzeczy, jak nauczanie maszynowe, filtrowanie bayesowskie, metody statystyczne, algorytmy genetyczne, sieci neuronowe, drzewa decyzyjne, za-

awansowane techniki optymalizacyjne. Z tego względu większość firm na świecie powierza budowę takich systemów wyspecjalizowanym zespołom naukowców. Dobrze rozumienie nauczania maszynowego i metod statystycznych będzie miało coraz większe znaczenie w wielu dziedzinach. Zbieranie i interpretacja ogromnej ilości danych tworzonych przez ludzi w systemach informatycznych pozwoli osiągać korzyści biznesowe w miejscach jeszcze nieodkrytych i niezagospodarowanych. ■

Marcin Majda, doktorant w Grupie Inteligentnych Systemów Informatycznych Katedry Informatyki Akademii Górniczo-Hutniczej w Krakowie. Specjalista od systemów rekomendacyjnych w Software Mind S.A. Jest doświadczonym software engineerem z certyfikatami SCJA, SCJP, SCWCD. marcin.majda@softwaremind.pl