Eksploracja masywnych danych
Projekt – Sprawozdanie II
Sformułowanie problemu i przetwarzanie danych

Tomasz Kuliński Wojciech Mioduszewski

31 stycznia 2013

1 Sformułowanie problemu

Konkurs nosi nazwę: "Personalized Web Search Challenge". Zakończył się 10 stycznia 2014, a jako nagrodę można było otrzymać 5000$, 3000$ lub 1000$ za odpowiednio pierwsze, drugie i trzecie miejsce. Link:

http://www.kaggle.com/c/yandex-personalized-web-search-challenge

Zadanie polega na tym aby uporządkować na nowo listę linków zwracanych przez przeglądarkę dla danego zapytania pod konkretnego użytkownika. W tym celu do dyspozycji otrzymujemy dane historyczne przeglądarki oraz obecnej sesji.

1.1 Opis metody oceny rozwiązań

Podstawowym problemem jest zakwalifikowanie każdego linka do jednej z trzech kategorii: **0 - nieważnych, 1 - ważnych i 2 - bardzo ważnych**. Do pierwszej powinny zaliczać się te linki, które nie zostaną wybrane przez użytkownika, albo wskazywane przez nie strony będą przez niego analizowane krócej niż 50 jednostek czasu. W grupie drugiej mają się znaleźć linki do tych stron, które zostaną wybrane przez użytkownika i przeglądane od 50 do 399 jednostek czasu. Wszystkie linki do stron przeglądanych przez przynajmniej 400 jednostek, lub te, które były ostatnimi akcjami w ramach danej sesji, mają się znaleźć w trzeciej grupie.

Zadaniem jest takie przeoragnizowanie kolejności url_id w zapytaniach testowych, aby były posortowane zgodnie z malejącą wartością kategorii tak, aby przedstawić użytkownikowi jak najbardziej stosowne linki. Oceną rozwiązania jest miara Normalized Discounted Cumulative Gain, wyrażająca się wzorem:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

gdzie: $DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i}-1}{\log_2(i+1)}$, a $IDCG_p$ jest idealną wartością miary.

1.2 Założone cele

Jak okazało się w czasie konkursu, odesłanie zapytań konkursowych do przeporządkowania jako gotowe rozwiązanie dawało bardzo wysoki wynik: 0.79133. Naturalnym celem jaki sobie zatem postawiliśmy, było przekroczenie tego rezultatu.

2 Opis metody predykcyjnej

Rozwiązanie bazowe, jakie zastosowaliśmy, to wyliczenie pewnego współczynnika dla wszystkich linków pokazywanych użytkownikowi. Wygląda on następująco:

$$url.statistic = \frac{2 * N_{VeryRelevant} + N_{Relevant}}{N_{Occurances}}$$

gdzie $N_{VeryRelevant}$ to liczba kliknięć zaklasyfikowanch jako bardzo ważne, $N_{Relevant}$ to liczba kliknięć zaklasyfikowanych jako ważne, a $N_{Occurances}$ to liczba wystąpień linków w wynikach wyszukiwania. Dalsze przetwarzanie to zmiana kolejności wyników zapytań testowych z częściowym uwzględnieniem początkowej kolejności. Dokładny wzór na współczynnik, po którym następowało sortwanie wygląda następująco:

$$final.importance = statistic.factor * url.statistic + (10 - poz)$$

gdzie url.statistic to ustalana przez nas waga, natomiast poz to domyślna pozycja linku w wynikach.

Naszym pomysłem było zebranie użytkowników w grupy i obliczanie wspomnianych współczynników dla każdej grupy z osobna. Przyjęliśmy dwa warianty - w jednym podawaliśmy konkretną liczbę grup (dokładnie 1/10000 liczby użytkowników) i następnie dodawaliśmy użytkowników do grupy z najbardziej podobnymi użytkownikami. Drugi wariant zakładał dodawanie użytkowników do grup tylko wtedy, gdy podobieństwo przekraczało pewien próg (0.5), w przeciwnym wypadku użytkownik tworzył nową grupę. W obu metodach porównywaliśmy użytkowników miarą Jaccarda, biorąc pod uwagę zbiory termów używanych przez nich w zapytaniach.

3 Wyniki eksperymentalne na zbiorze treningowym

Czasy uczenia się i czas obliczeń:

Metoda	Czas uczenia	Czas obliczeń
Rozwiązanie bazowe	12min	5s
Grupowanie metoda 1	3h 20min	2min 30s
Grupowanie metoda 2	2h 30min	8min

Table 1: Czasy wykonywania rozwiązań

W pierwszym podejściu do grup (mało grup) zarówno czas uczenia, jak i produkowania ostatecznego rozwiązania był nieco mniejszy w porównaniu do rozwiązania drugiego. Gdy liczyliśmy drugim sposobem, czas poświęcony na uczenie został przez nas mocno ograniczony (wzięliśmy pod uwagę jedynie najliczniejszych 20% grup, warto jednak zaznaczyć, że dalsze grupy miały znacznie mniejsze liczności, a ponad połowa grup była jednoosobowa), dzięki czemu czas uczenia wydaje się akceptowalny, jednak zwracanie ostatecznego rozwiązania było wydłużone (i to pomimo tak dużego zredukowania ilości grup). W ostatecznych wynikach metoda 2 okazała się być skuteczniejszą, jednak decydowanie o jej poprawności na podstawie wyniku który i tak jest mniejszy od wyniku początkowego nie jest do końca bezpieczne.

4 Wynik końcowy

Metoda	Wynik
Rozwiązanie początkowe	0.78881
Rozwiązanie bazowe	0.79133
Grupowanie metoda 1	0.78495
Grupowanie metoda 2	0.79033

Table 2: Wyniki różnych metod

Ostateczny wynik jaki uzyskaliśmy używając metod predykcyjnych, niestety nie przekroczył rozwiązania początkowego, korzystającego z istniejącego rankingu Yandexa. Co za tym idzie - ostatateczny wynik w serwisie Kaggle wyniósł 0.79133, gdzie razem z osiemdziesięcioma innymi drużynami dzieliliśmy 93 miejsce. Warto wspomnieć, że najlepsza drużyna (która jak się okazało była drużyną wystawioną przez Yandex) uzyskała wynik: 0.80725, który przebija rozwiązanie domyślne zaledwie o 0.01592. Liczby te pokazują, że ulepszenie istniejącego już uporządkowania było nietrywialne. Jeśli chodzi o nasze rozwiązanie, które wiązało się przeporządkowaniem istotnej ilości danych, to było ono gorsze od początkowego zaledwie o 0.001 - jednak ciężko powiedzieć czy było ono na tyle dobre, czy jednak zmieniało tylko na tyle mało istotne urle (zamiana urli o tej samej ocenie nie zmienia wyniku), a co za tym idzie, miało mało okazji by się "pomylić".

Wydaje się, że przyjęte przez nas metody są znacznie mniej złożone od tych, które są domyślnie stosowane w Yandex'ie i w związku z tym zmiana początkowego rozwiązania psuje jego całkiem dobre uporządkowanie. Całkiem możliwe, że winowajcą są niewystarczające algorytmy podziału użytkowników na grupy. Próbowaliśmy zastosować algorytm KMeans, jednak po pierwsze przy tak dużej ilości użytkowników działał on stanowczo zbyt długo (pojedyncza iteracja przekraczałaby dobę), a po drugie nie mieliśmy opracowanej dobrej metody wyznaczania elementu średniego dla grupy użytkowników.

5 Podsumowanie

Uczestnictwo w omawianym konkursie było ciekawym doświadczeniem i wiele nas nauczyło - nie tylko z dziedziny eksploracji danych czy uczenia maszynowego. Mimo, że wysoce sobie to cenimy, to po tych zawodach czujemy pewien niesmak, gdyż szczere zaangażowanie i duża ilość pracy nie została przełożona na jakikolwiek wynik - wszytkie nasze predykcje spełzły na niczym, wobec rozwiązania początkowego, a na początku założony cel pozostał tylko zbyt wysoko postawioną poprzeczką.