

Opis scenariusza

- WVCorp: firma zatrudniająca Ciebie (analityka danych).
 - Firma WVCorp dysponuje forami tematycznymi i dyskusyjnymi dla każdego produktu, gdzie klienci mogą dyskutować o problemach i cechach.
 - “Szum medialny” — wtedy, gdy jakiś temat na forum użytkowników ma bardzo duży poziom aktywności – co wskazuje na zainteresowanie użytkowników tym zagadnieniem.
- eRead — czytnik e-booków wyprodukowany przez firmę WVCorp.
- TimeWrangler — aplikacja do zarządzania czasem wyprodukowana przez firmę WVCorp.
- BookBits — czytnik e-booków wyprodukowany przez konkurencję.
- GCal — stworzona przez inną firmę rozproszona usługa kalendarzowa, którą można zintegrować z aplikacją TimeWrangler.

Uwagi opisujące fikcyjny świat, w którym ma miejsce ta prezentacja.

Przewidywanie szumu medialnego na forach użytkowników

Hipotetyczna prezentacja
dla partnerów
przygotowana przez zespół
analityków danych WVCorp

Szum jest informacją

- Szum medialny — tematy na forum użytkownika o dużej aktywności; tematy, którymi użytkownicy są zainteresowani.
 - Cechy wyczekiwane przez klientów.
 - Istniejące cechy, które sprawiają problemy.
 - Trwały szum — rzeczywiste, trwające potrzeby klientów.
 - Problemy ani chwilowe, ani wynikające z trendów.
- **Cel: przewidywanie tematów mających trwały szum medialny.**

Partnerzy są zasadniczo bardziej zainteresowani zadaniem predykcyjnym i akceptują umiarkowaną motywację („szum medialny jest przydatny”). Jednak w przypadku pewnych grup odbiorców (zwłaszcza znajdujących się poza Twoją organizacją) możesz chcieć nakreślić problem biznesowy (1–2 pierwsze slajdy w prezentacji przeznaczonej dla sponsora projektu) w celu wyznaczenia kontekstu.

Powiązana praca

- *Predicting Movie Success and Academy Awards Through Sentiment and Social Network Analysis*
 - Krauss, Nann i in. *European Conference on Information Systems*, 2008
- Fora IMDB, strona Box Office Mojo.
- Zmienne: intensywność dyskusji, pozytywność.
- *Predicting Asset Value through Twitter Buzz*
 - Zhang, Fuehres, Gloor, *Advances in Collective Intelligence*, 2011
- Analiza szeregów czasowych na wybranych słowach kluczowych.

W prezentacjach dla pracowników akademickich zazwyczaj wstawiamy sekcję poświęconą powiązanej pracy: informacje na temat innych osób, które mogły prowadzić badania dotyczące podobnych problemów, podobieństwa w przyjętej strategii badawczej, a także różnice (a być może także powód, dla którego wybrałeś inną strategię).

Badanie pilotażowe

- Zebrano na forum dane z trzech tygodni:
 - 7900 tematów, 96 zmiennych,
 - 791 tematów odłożonych do oceny modelu.
 - 22% tematów z pierwszego tygodnia „szumiało” w drugim i trzecim tygodniu.
- Na danych z pierwszego tygodnia wyuczono model lasu losowego w celu określenia, które tematy będą „szumieć” w drugim i trzecim tygodniu.
 - Szum = utrzymywany wzrost 500+ aktywnych dyskusji na dany temat w ciągu dnia w porównaniu z pierwszym dniem w pierwszym tygodniu.
- Opinie od pięcioosobowego zespołu menedżerów produktu; jak przydatne były wyniki?

Szczegółowe wprowadzenie do naszych działań. Pozostaw punkt dotyczący menedżerów produktu po to, aby zachować „istotność” modelu w dyskusji.

Zmienne modelu

- Użyliśmy najpierw wskaźników już monitorowanych przez system.
 - Liczba autorów na temat.
 - Liczba dyskusji na temat.
 - Liczba wyświetleń tematu użytkownikom forum.
 - Średnia liczba osób biorących udział w dyskusji na dany temat.
 - Średnia długość dyskusji na dany temat.
 - Częstość przesyłania dyskusji na dany temat do mediów społecznościowych.
- Jeden z problemów — jedynie wskaźniki punktowe.
 - W idealnym przypadku chcemy mierzyć ewolucję.
 - Np. czy liczba autorów maleje/wzrasta? Jak szybko?
 - Analiza szeregów czasowych.
 - Jak dobrze możemy sobie radzić z tym, co mamy?

Tej grupie odbiorców przekazujemy mnóstwo szczegółów.

Model lasu losowego

- Wydajny w przypadku dużej liczby danych i zmiennych wejściowych.
- Kilka ogólnych założeń dotyczących rozkładu zmiennych/interakcji.
- Ograniczyliśmy złożoność, aby zminimalizować przetrenowanie.
 - Maksymalnie 100 węzłów na drzewo.
 - Minimalny rozmiar węzła: 20.
- Większa ilość danych wyeliminuje konieczność stosowania tych ograniczeń.

Szczegóły dotyczące wybranej strategii modelowania.

Wyniki

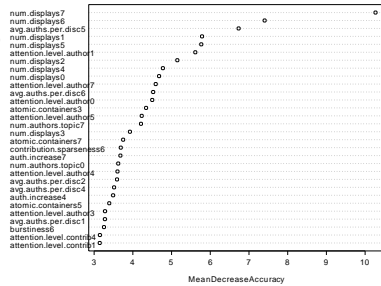
- 84% pełności, 83% precyzji.
- Ponad czterokrotnie zmniejszone ręczne przeglądanie forów.
 - Zmniejszona liczba tematów do analizy: ze 791 do 184.
- Użytkownicy: 75% rozpoznanych tematów dało "wartościową opinię".

	Przewid. brak szumu	Przewid. szum	
Brak szumu	579	35	614
Szum	28	149	177
Łącznie	607	184	791

Wydajność modelu i inne powiązane fakty.

Istotność zmiennych

- Kluczowe dane wejściowe:
 - Liczba wyświetleń danego tematu użytkownikowi (num.displays)
 - Liczba autorów biorących udział w dyskusji (attention.level.author)
- Zmienne prędkości (ang. velocity variable) dla tych dwóch danych wejściowych mogą usprawnić model.



Dyskusja na temat zmiennych wejściowych i ich wpływu na model.

Przykładowe odkrycie

- Temat: TimeWrangler → Integracja z GCal.
 - Liczba dyskusji wzrosła od czasu opublikowania GCal v7.
 - Zdarzenia GCal niespójnie wykrywane; niewłaściwe oznakowanie.
 - Zadania TimeWrangler przechodzą do złej aplikacji Gcalendar.
 - **„Gorąco” na forach, jeszcze zanim zrobiło się „gorąco” w dziennikach zdarzeń obsługi klienta.**
 - Aktywność na forum pobudziła model dwa dni po aktualizacji GCal.
 - Obsługa klienta spóźniona o tydzień.

Przykładowy rezultat.

Plany na przyszłość

- Lepsze zmienne wejściowe.
 - Zmienne wymiarów i prędkości.
 - Szybkość wzrostu/spadku liczby autorów.
 - Szybkość wzrostu/spadku wyświetleń liczby tematów.
 - Informacje o nowych osobach odwiedzających forum.
 - Jakie pytania zadają nowi użytkownicy?
- Badanie optymalnego harmonogramu ponownego uczenia modelu.

Dziękuję