FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Evaluace algoritmů lokálně senzitivního hashování (LSH) v doporučovacích systémech

Ladislav Martínek

Ročník: 3
Obor: Znalostní inženýrství

Katedra aplikované matematiky
Vedoucí práce: Ing. Tomáš Řehořek

Klíčová slova: aproximace algoritmu nejbližších sousedů, lokálně senzitivní hashování, doporučovací systémy, implementace testovací knihovny, experimenty a analýza výsledků

14. dubna 2018

1 Úvod

V dnešní moderní společnosti, kde neustále roste objem dat nabízený na internetu, je pro uživatele velmi těžké vyhledávat informace nebo produkty. Proto se společnosti zabývají přizpůsobováním obsahu každému uživateli na internetu např. doporučením, co by se uživateli mohlo líbit. Při doporučovaní se snaží nabídnout např. film, článek, který by mohl uživatele zajímat. Doporučování na internetu je nutné provádět v reálném čase, což se současnými objemy dat je velice problémové. Z tohoto důvodu je výhodné, zároveň i nutné se zabývat zrychlováním doporučovacích systému.

Existuje mnoho metod, které se zrychlováním zabývají. V práci se věnuji jedné konkrétní metodě a to lokálně senzitivnímu hashování, pomocí kterého lze dosahovat několikanásobného zrychlení výpočtu při malé ztrátě přesnosti. Výstup mé práce by mel pomoci programátorům s testováním a vyhodnocováním různých parametrizací použité metody pro zrychlení vyhodnocování doporučovacích systémů.

Téma jsem si zvolil, neboť doporučovací systémy jsou využívány mnoha společnostmi na internetových stránkách, kde nám jejich přítomnost usnadňuje hledání informací. Nicméně je velmi výpočetně náročné vyhodnocovat všechna nasbíraná data v reálném čase, proto se v práci zabývám aproximací doporučovacích systémů, konkrétně pak návrhem, implementací a následným vyhodnocením knihovny na aproximaci hledání nejbližších sousedu.

2 Cíl práce

Cílem rešeršní části práce je získat přehled o doporučovacích systémech, seznámit se s algoritmy nejbližších sousedu a nastudovat algoritmy lokálně senzitivního hashování a možnosti využití těchto metod při aproximaci algoritmu nejbližších sousedu. Dalším navazující cíl rešeršní části je představení metod vyhodnocování úspěšnosti představených algoritmů a doporučovacích systému založených na těchto algoritmech.

Cílem praktické části práce je navrhnout způsob aplikace metod lokálně senzitivního hashování za účelem aproximace algoritmu hledání nejbližších sousedu, poté návrh a implementace frameworku pro testování úspěšnosti algoritmu lokálně senzitivního hashování v doporučovacích systémech. V neposlední řadě je cílem otestování implementace algoritmu s různou parametrizací na dvou datových sadách a diskuze výsledku, konkrétně poměr zrychlení výpočtu na úkor zhoršení přesnosti.

3 Analýza

Nejprve stručně vymezím pojmy a podle literatury popíši přístupy, kterým se v práci věnuji. Na to naváži analýzou současného stavu řešení aktuální problematiky.

3.1 Základní pojmy

Doporučovací systémy jsou nástroje a techniky, které generují doporučení položek, které by mohly být zajímavé nebo užitečné pro uživatele [1, pg.1]. Doporučením mohou být například produkty k prodeji, filmy nebo i slevové kupóny od obchodních řetězců.

Základním přístupem v doporučovacích systémech je filtrování obsahu a informací pomocí spolupráce mezi více uživateli (Kolaborativní filtování, angl. Collaborative filtering). V mé práci se budu věnovat filtrování pomocí spolupráce uživatelů. Klíčovou myšlenkou je, že pokud uživatelé sdílejí zájem o určité položky, budou tento zájem sdílet i nadále a budou preferovat podobné položky [4].

Algoritmy nejbližších sousedů (angl. k-nearest neighbors, k-NN) jsou přístupem kolaborativního filtrování a vždy fungují v konkrétním měřitelném prostoru, který muže být definován různými způsoby. Existence metriky je důležitá pro porovnávání podobnosti nebo vzdálenosti jednotlivých sousedů.

Lokálně senzitivní hashování (angl. Locality sensitive hashing, LSH) je metoda aproximace algoritmu k-NN. Podobně jako i jiné aproximační metody je LSH randomizované a náhodnost je typicky využita ke konstrukci datové struktury pro vytvoření modelu [13].

3.2 Současný stav

Co se řešilo...

4 Realizace

Jak řeším...

5 Závěr

Cílem práce byl návrh a implementace knihovny pro testování metod lokálně senzitivního hashování, úspěšnosti těchto metod a jejich různých parametrizací.

Knihovna implementována v rámci práce umožňuje testovat různé parametrizace a modely LSH na různých databázích, které lze specifikovat v konfiguračním souboru. Z výsledných dat generovaných frameworkem je možné pomocí nezávislého modulu generovat grafy pro lepší prezentaci výsledků.

Dosažené výsledky při testování různých parametrizací jsou velice uspokojivé. Na dvou otestovaných datových sadách bylo zjištěno, že lze dosahovat času kolem 5 %. Úspěšnost doporučování se při této aproximaci pohybuje okolo 97 - 99 %. Při této úspěšnosti dosahují modely LSH přibližně o 20 % lepší catalog coverage vzhledem k referenčnímu modelu. Doporučovací systém s metodami LSH tedy doporučuje větší množství různých položek, proto mohou být metody LSH užitečnější tím, že kromě časového zrychlení, mohou doporučovat větší část položek z databáze.

Zajímavým poznatkem je zjištěná závislost mezi přesností na 10 nejbližších sousedů a úspěšnosti doporučování. Díky tomuto zjištění bylo možné otestovat velké množství různých parametrizací, protože testování přesnosti na 10 nejbližších sousedů je efektivnější. Testováním byli zjištěny některé hranice nastavení jednotlivých modelů a také nastavení pomocí kterého lze dosahovat optimálních výsledků.

V neposlední řadě je zajímavý poznatek nastavení parametrů testu pro testování metod LSH pomocí frameworku. Pro každou databázi je nutné zvolit požadované parametry při testování z ohledem na velikost databáze. Tyto parametry lze podle počtu uživatelů lehce dopočítávat a tím přizpůsobit testování pro konkrétní databázi.

V budoucnu by mohli tyto metody implementované jako součást doporučovacího systému, kde by bylo nutné se zabývat jejich paralelizací. Metody LSH lze velice efektivně paralelizovat obdobně jako samotný algoritmus k-NN. V budoucí práci by bylo také určitě nutné omezit maximální počet uživatelů v jednom bucketu.

Při využití v online doporučovacím systému by bylo nutné se zabývat automatizací nastavení modelu, tedy zvoleným množstvím hashovacích funkcí, množství bucketů a velikost okolí procházených bucketů. Zároveň na online datech není znám předem přesný objem dat ani počet uživatelů, proto by bylo zajímavé navrhnout způsob, kterým by bylo možné přizpůsobovat model při online běhu (např. přehashováním, přidávání nebo odebírání hashovacích funkcí).

6 Poznámky

1. Úvodní strana - název, jméno autora, jméno vedoucího práce, ročník a obor studia, případně logo fakulty

2. Klíčová slova - cca 5 až 10 odborných termínů charakterizujících zaměření bakalářské práce

3. Úvod
4. Definice cíle(ů) bakalářské práce
5. Analýza současného stavu řešení problému - kdo a jakým způsobem již problém řešil, k čemu došel, výhody a nevýhody řešní
6. Aktuální stav řešení - řešení, které student zvolil
7. Závěr - míra splnění cíle(ů)
8. Seznam literárních zdrojů dle ČSN ISO 690

test[1]

Reference

[1] O'rourke, J.: *Art gallery theorems and algorithms*, ročník 57. Oxford University Press Oxford, 1987.