Od klíčových slov k významu: Jak vektorové indexy mění Vyhledávání na Seznamu

24.9.2024 - Lenka Lasoňová | Seznam.cz

Vyhledávání informací na internetu se stalo nedílnou součástí našich životů. Ať už potřebujeme rychle zjistit, jak opravit zaseknutý zip, najít nejlepší kavárnu v okolí nebo vybrat dárek k narozeninám – internetové vyhledávače jsou často naší první zastávkou. A s rostoucí sofistikovaností technologií se mění i způsob, jakým s vyhledávači komunikujeme.

Časy se mění. Zatímco dříve jsme zadávali krátké, úderné fráze složené z klíčových slov, dnes stále častěji pokládáme dotazy v přirozeném jazyce. Místo „bolest hlavy“ se ptáme „Co pomáhá na migrénu?“ nebo „Jak rychle ulevit od bolesti v krku?“. Tento posun klade na vyhledávače nové nároky – musí porozumět nejen jednotlivým slovům, ale i kontextu a záměru celé věty.

Na počátku internetového vyhledávání stály invertované indexy

Tento systém funguje na principu mapování slov na dokumenty, kdy pojem dokument označuje webovou stránku nebo její část, ve které se vyskytují. Zjednodušeně si můžete představit obrovskou tabulku, kde v jednom sloupci jsou všechna slova z indexovaných dokumentů a v druhém sloupci seznamy odkazů na dokumenty, kde se tato slova nacházejí. Když uživatel zadal dotaz, vyhledávač jednoduše našel v této tabulce odpovídající slova a vrátil seznam dokumentů, kde se tato slova vyskytovala. Tento systém byl a je velmi efektivní pro rychlé vyhledávání na základě klíčových slov.

Invertované indexy však mají své limity, zejména v oblasti sémantiky. Bez rozsáhlých slovníků synonym a příbuzných slov nedokáží rozpoznat, že „auto“ a „vozidlo“ mohou v určitém kontextu znamenat totéž nebo že „doktor“ a „lékař“ označují stejnou profesi. Chybí jim porozumění významu a kontextu slov.

Vektorová reprezentace jazyka: Cesta k porozumění významu

Průlom v této oblasti přišel s technikami vektorové reprezentace jazyka. Tato metoda převádí slova, fráze nebo celé dokumenty na číselné vektory v mnohorozměrném prostoru. V tomto prostoru jsou slova s podobným významem blízko u sebe.

Klasickým příkladem je vztah mezi slovy „král“, „královna“, „muž“ a „žena“. Pokud od vektoru reprezentujícího slovo „král“ odečteme vektor „muž“ a přičteme vektor „žena“, dostaneme se velmi blízko k vektoru reprezentujícímu slovo „královna“. Tento jednoduchý příklad ukazuje, jak vektorové reprezentace zachycují významové vztahy mezi slovy.

Metody vektorové reprezentace textu prošly za poslední dekádu zásadním vývojem. Významným okamžikem bylo představení techniky Word2Vec v roce 2013. Ta umožnila efektivní vytváření vektorových reprezentací slov s ohledem na sémantické vztahy.

Následovaly metody jako GloVe nebo ELMo, které přinesly další zlepšení, ale zásadní průlom nastal v roce 2018 s příchodem jazykových modelů založených na architektuře transformerů (např. BERT).

Tyto modely vytvářejí hluboké, kontextově závislé reprezentace nejen slov, ale i celých vět a dokumentů. Nejnovější generace modelů, jako je GPT-3, dokáže navíc generovat vysoce kvalitní vektorové reprezentace pro široké spektrum jazykových úloh.

Vektorové indexy jsou revoluce ve vyhledávání

Zatímco vektorová reprezentace jazyka umožnila porozumění významu slov, vektorové indexy přinášejí efektivní způsob, jak v těchto vektorech vyhledávat. Fungují na principu organizace vektorových reprezentací do speciálních datových struktur, které umožňují rychlé nalezení nejbližších sousedů v mnohorozměrném prostoru.

Tyto struktury, často založené na hierarchických stromech nebo grafech, dokáží efektivně prohledávat miliardy dokumentů na internetu tím, že rychle identifikují skupiny podobných vektorů. Díky tomu mohou vektorové indexy objevit relevantní dokumenty, které by tradiční vyhledávače založené na klíčových slovech mohly přehlédnout.

Představte si například situaci, kdy hledáte rady na „rychlou večeři pro děti“. Dokument, který mluví o „jednoduchých jídlech pro rodiny“ nebo „snadných receptech pro zaneprázdněné rodiče“, by tradiční vyhledávač pravděpodobně přehlédl.

Vektorové indexy však dokáží rozpoznat sémantickou podobnost těchto frází a přinést relevantní výsledky, i když neobsahují přesná slova z daného dotazu. Kromě toho si dokáží poradit s náročnými překlepy a v některých případech i s fonetickými přepisy. Bez této schopnosti by mnoho dokumentů zůstalo prakticky nedohledatelných, což ukazuje, jak zásadní jsou vektorové indexy pro moderní vyhledávání.

Přes nesporné výhody vektorových indexů čelí jejich implementace výzvám v podobě vysoké výpočetní náročnosti, nároků na paměť a občasné nepřesnosti při vyhledávání specifických informací, jako jsou přesná jména, data nebo lokality. Proto se v praxi osvědčuje kombinace vektorových indexů s tradičními invertovanými indexy, což umožňuje využít silné stránky obou přístupů – sémantické porozumění vektorových indexů a přesnost invertovaných indexů při vyhledávání konkrétních termínů.

Zapojení vektorových indexů do Vyhledávání na Seznamu

Seznam aktivně využívá vektorové indexy ve Vyhledávání od roku 2020. Pro tento účel využíváme vlastní implementaci, která nám umožňuje plnou kontrolu a optimalizaci celého procesu. Naše řešení je postaveno na populární knihovně *hnswlib* , která využívá grafovou strukturu a je známá svou efektivitou, škálovatelností a rychlostí při práci s vysokodimenzionálními daty.

Zpočátku jsme pro převod dokumentů do vektorových reprezentací používali modely typu *fastText,* následně deriváty modelu BERT založené na předtrénování modelu ELECTRA. Postupným vývojem jsme však přešli na jazykové modely předtrénované RetroMAE přístupem, které nám pomáhají ještě lépe zachytit význam textu. Naše modely jsou speciálně trénované na českém jazyce a obsahu, díky čemuž dokáží porozumět jemným nuancím a specifikům češtiny. Pokud vás zajímají naše nejnovější modely z tohoto roku, doporučujeme přečíst si články tady a zde.

Při vytváření vektorových reprezentací dokumentů bereme v úvahu více faktorů, abychom získali co nejkomplexnější reprezentaci obsahu. Jako vstupy do našich modelů používáme například:

- Často obsahuje klíčové informace o obsahu.

- **URL adresu**
  Může poskytnout dodatečné kontextové informace.
- **Odstavce z dokumentu**
  Analyzujeme samotný obsah pro hlubší porozumění tématu.

V současnosti používáme dva modely pro generování vektorových reprezentací:

1. Model ve velikosti base

Zpracovává titulek, URL a počátek dokumentu, produkuje 256dimenzionální vektorové reprezentace.

2. Model ve velikosti small

Zaměřuje se na jednotlivé odstavce dokumentu a produkuje 128dimenzionální vektorové reprezentace. Tato kombinace nám umožňuje zachytit jak celkový kontext dokumentu, tak i detailní informace z jeho obsahu.

Tyto modely jsou teď základem pro naši implementaci vektorových indexů ve Vyhledávání. Jsou ale stále jen jedním z mnoha dílků, které nám umožňují lépe porozumět uživatelským dotazům a poskytovat relevantnější odpovědi. Vývoj vyhledávače je kontinuální proces, v němž vektorové indexy představují důležitý, ale zdaleka ne poslední krok na cestě k dokonalejšímu vyhledávání.

https://blog.seznam.cz/2024/09/od-klicovych-slov-k-vyznamu-jak-vektorove-indexy-meni-vyhledavani-na-seznamu