Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Katedra softwarového inženýrství, Fakulta informačních technologií,
České vysoké učení technické v Praze

VYHLEDÁVÁNÍ NA WEBU A V MULTIMEDIÁLNÍCH DB (BI-VWM)

Projekt I - 3

INFORMATION RETRIEVAL – VEKTOROVÝ MODEL

RNDr. David Hoksza, PhD.

ZADÁNÍ

Cílem projektu je implementace vektorového systému ukládání dat (tj. poreprocessing a indexování) spolu s možností dotazování z GUI.

VSTUP

Dotaz – seznam termů spolu s váhami.

VÝSTUP

Seznam databázových dokumentů odpovídající dotazu v klesajícím pořadí podobnosti.

INFORMACE/POTŘEBNÉ ZNALOSTI

Vektorový model je jeden ze způsobů jak prohledávat kolekci dokumentů adresující nedostatky boolovského modelu. Oproti boolovskému modelu není binární, tj. pro každý term neobsahuje informaci o tom, ve kterém dokumentu se daný term vyskytuje, nýbrž informaci o tom, *jak moc* se ve kterém dokumentu vyskytuje. Dotaz je pak tvořen seznamem termů a váhami (důrazem), které uživatel jednotlivým termům přiřazuje.

Dotaz je vyhodnocován oproti kolekci dokumentů, tj. každý dokument lze chápat jako objekt databáze. Nejjednodušší možností přístupu k této databázi je procházet každý dokument zvlášť a dotazovat se, jak moc je daný dokument podobný dotazu. S rostoucí velikostí kolekce je ovšem takovýto přístup nevyhovující a proto je třeba ukládat data ve formě vhodné pro vyhledávání. Stejně jako u boolovského modelu projde každý dokument nejdříve fází preprocesingu, kdy jsou z dokumentu odstraněna nevýznamová slova (tj. slova, která se nesou málo informace, jako např. spojky a předložky) a významová slova jsou “stemmovány” (jednodušší proces) nebo “lematizovány” (sofistikovanější proces) za účelem získání základů slov.

Po preprocesingu máme tedy k dispozici kolekci slov, kterou je třeba uložit takovým způsobem, aby v ní šlo efektivně vyhledávat. U boolovského modelu je každý dokument uložen jako binární vektor, čímž dostáváme tzv. term-by-document matici, kde na *i*-tém řádku v *j*-tém sloupci je 1, právě tehdy pokud je term *i* obsažen v dokumentu *j*. Takový přístup nedokáže rozlišit, jak moc daný term vystihuje dokument, v kterém se nachází. Nelze říci, zda se term *i* vyskytuje v dokumentu pouze okrajově, nebo je celý dokument právě o tomto termu. Z toho důvodu jsou v term-by-document matici reálné hodnoty v rozmezí 0 až 1, definující váhu (důležitost) termu pro dokument. Určování vah je typicky založeno na frekvenci výskytu termu v dokumentu a výskytu termu přes celou kolekci. Nejznámější schéma založené na tomto principu se nazývá *tf-idf* (term frequency - inverse document frequency) schéma. Každý dokument je pak možné popsat *n*-dimenionálním vektorem (*n* je velikost slovníku) a tedy lze chápat jako bod v n-dimenzionálním prostoru.

Stejně jako dokument lze reprezentovat i dotaz. Dimenze odpovídající termům, které zadal uživatel, mají hodnoty určeny dotazem, ostatní dimenze mají hodnotu 0. Definujeme-li pak nějakou vzdálenost mezi dvěma body v *n*-dimenzionálním prostoru (u vektorového modelu je to kosinová vzdálenost), lze podobnost dokumentů chápat jako převrácenou vzdálenost bodů, které je reprezentují.

Podobně jako u boolovského modelu lze pro efektivní implementaci využít invertované seznamy. Tyto slouží ve vektorovém modelu k identifikaci dokumentů obsahujících dané termy (jako u boolovského modelu). Identifikované dokumenty jsou pak setříděny podle podobnosti k dotazu (viz předchozí odstavec).

Vektorý model se tedy skládá z následujících částí:

1. Extrakce a preprocesing termů z dokumentů.
2. Efektivní uložení dokumentů v datové struktuře (invertovaný seznam).
3. Vyhodnocovací/dotazovací modul využívající strukturu z předchozího kroku.

STAVBA APLIKACE

Aplikace by měla obsahovat:

- Extrakce termů.
- Identifikace nevýznamových slov.
- Stemming/lematizace.
- Výpočet vah termů.
- Implementace indexovací struktury.
- Vyhodnocení dotazu oproti indexovací struktuře.
- Webový interface (zadání dotazu a vizualizace výsledku).

POZNÁMKY K ŘEŠENÍ

V rámci projektu je třeba implementovat jak vektorový model umožňující neprocházet celou kolekci (invertovaný seznam například), tak sekvenční průchod, tj. procházení kolekce dokumentů bez využití indexu. Sekvenční průchod je pak možné použít k porovnání výsledků vyhledávání vzhledem k vektorovému modelu.

Lze využít knihovny na parsování dokumentů, příp. preprocessing.

DATA

Datová sada by měla obsahovat alespoň tolik dokumentů, aby bylo možné pozorovat výhody použití vektorového modelu oproti sekvenčnímu průchodu. Zdroj dat je libovolný – např. offline verze nějakého webového serveru (novinové články, …).

EXPERIMENTY

V tomto projektu lze mimo jiné provádět srovnání vektorového modelu se sekvenčním průchodem s ohledem na čas vykonání dotazu. Lze také testovat vliv různých vnitřních parametrů na výkon algoritmu (např. různé nastavení v invertovaném seznamu) apod.

ZDROJE

- Přednáška *Vyhledávání textu - Booleovské modely. Implementace.*
- Přednáška *Vyhledávání textu - Vektorové modely. Implementace.*
- Jaroslav Pokorný, Václav Snášel, Dušan Húsek. *Dokumentografické Informační Systémy*. Karolinum, 1998