Masarykova univerzita
Fakulta informatiky

Implementace fulltextového vyhledávání v issue tracking systému

Bakalářská práce

Jiří Holuša

Brno, Jaro 2014

Prohlášení

Prohlašuji, že tato bakalářská práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Jiří Holuša

Vedoucí práce: Mgr. Filip Nguyen

Poděkování

TODO: poděkování

Shrnutí

TODO: abstrakt

Klíčová slova

TODO: klíčová slova

Obsah

1 Úvod

Úvod

2 Vyhledávání

Tato kapitola stručně popisuje způsob vyhledávání v nejčastějším datovém úložišti - relačních databázích - a uvádí jeho nedostatky. Poté se detailněji věnuje jednou z možností jejich řešení, a to fulltextovým vyhledáváním. Uvádí nezbytnou teorii k pochopení principů, jak fulltextové vyhledávání funguje, jeho výhody a nevýhody.

2.1 Vyhledávání v relačních databázích

Relační databáze poskytují vysoce výkonný přístup k datům a široké možnosti pro jejich správu. Díky svých schopnostem se staly nejpoužívanější technologií pro datové uložiště. Vyhledávat v datech lze přitom pouze dvěma způsoby: porovnání obsahu buňky a operátor LIKE.

Porovnání obsahu buňky funguje na velice jednoduchém principy úplné shody obsahu. V následujícím příkladu vidíme dotaz v jazyce SQL, který vybere právě ty záznamy z tabulky People, které mají hodnotu atributu name rovnou "Bruce Banner". SELECT * FROM People WHERE name = 'Bruce Banner'

Nebudou tedy vybrány žádné jiné záznamy, přestože by obsah atributu name měly např. "Bruce Banners"či dokonce ani "Bruce Banner "(přebytečná mezera na konci). Výhodou tohoto řešení je efektivita a jednoduchost - jedinná nutná operace je pouze porovnání dvou řetězců, žádné dodatečné zpracování není potřeba.

Trochu více sofistikovaným způsobem je operator LIKE, který umožňuje (v omezené míře) používat pattern matching - vyhledávání pomocí vzoru. Podporovány jsou tzv. zástupné symboly, jež mohou mít v tomto kontextu jiný význam než jen právě daný znak, např. symbol % (procento) zastupuje libovolnou sekvenci znaků (třeba i žádnou) nebo znak . (tečka) libovolný, ale právě jeden znak. Níže vidíme příklad SQL dotazu, jenž nám vrátí všechny záznamy z tabulky People, které jejich jméno končí na "Banner". SELECT * FROM People WHERE name LIKE '

Nyní již dokážeme tímto dotazem získat jak lidi se jménem "Bruce Banner", tak i "Richard Banner".

2.2 Problémy vyhledávání v relačních databázích

V předchozí kapitole jsme si představili základní způsoby vyhledávání v relačních databázích. Nyní se podíváme na případy, kde nám tyto způsoby

nestačí nebo si s danou situací nedokáží poradit buď vůbec, nebo pouze neefektivně.

2.2.1 Vyhledávání přes několik tabulek

2.2.2 Vyhledávání jednotlivých slov

2.2.3 Filtrace šumu

2.2.4 Vyhledávání příbuzných slov

2.2.5 Oprava překlepů

2.2.6 Relevance

2.3 Fulltextové vyhledávání

3 Dostupné technologie

3.1 Apache Lucene

3.2 Hibernate Search

3.3 Elasticsearch

4 Implementace

Implementační část

5 Závěr

Závěr

Literatura

[1] Douglas Adams. *The Restaurant at the End of the Universe*. The Hitchhiker's Guide to the Galaxy. Pan Macmillan, 1980.

[2] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, 1 edition, 1994.