Systém pre vyhľadávanie chemických štruktúr

Autor: Ivan Ševčík Vedúci: Zbyněk Křivka

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Problém a motivácia

- Databázy chemických látok obsahujú v dnešnej dobe desiatky až stovky miliónov záznamov
- V databázach je potrebné rýchlo a efektívne vyhľadávať
  - Identifikácia látok odvodených v rámci simulácie
  - Nájdenie toxických látok alebo nových liečiv na základe hľadania zhodnej podštruktúry
- Vyhľadávanie ďalej komplikuje algoritmicky náročná, grafová reprezentácia chemických štruktúr
- Existujúce dostupné riešenia neposkytujú dostatočnú rýchlosť a škálovateľnosť

Popis riešenia

V rámci práce bol navrhnutý a implementovaný systém podporujúci identické a podštruktúrne vyhľadávanie. Modulárna architektúra systému (obr. 1) umožňuje jeho ľahké rozšírenie o ďalšie výpočetné prostriedky. Vďaka porovnaniu niekoľkých algoritmov pre hľadanie izomorfného podgrafu (obr. 2), ktoré tvoria jadro podštruktúrneho vyhľadávania, bolo možné zaužívaný algoritmus VF2 nahradiť algoritmom RI a výrazne tak skrátiť čas potrebný na dokončenie operácie. Dôležitým konceptom sú tiež molekulárne odtlačky (obr. 3), ktoré umožňujú obmedziť množinu prehľadávaných záznamov. Vzhľadom na požiadavku veľmi rýchleho identického vyhľadávania bola pre túto operáciu použitá reprezentácia chemických štruktúr pomocou identifikátora InChI. Výsledné riešenie je sprístupnené užívateľom formou webovej stránky podporujúcej zakreslenie vzoru pomocou štruktúrneho editora a zobrazenie výsledkov.

Zhodnotenie

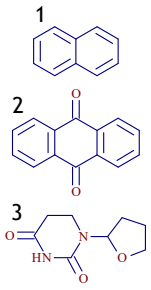
- V rámci práce sa podarilo splniť náročné požiadavky kladené na rýchlosť vyhľadávania
- Dokončením indexu pre odtlačky spomenutého v závere práce sa dosiahlo ďalšieho významného zrýchlenia
- Systém bude ďalej integrovaný ako súčasť väčšieho projektu zameraného na prácu a hľadanie súvislostí v databázach malých molekúl

Výsledné merania

Čas potrebný na dokončenie operácie vyhľadávania v databáze naplnenej desiatimi miliónmi záznamov.

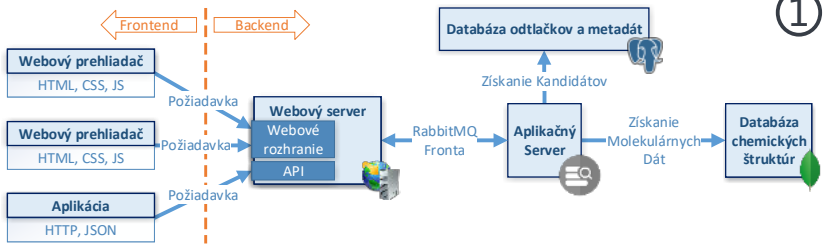
Vyhľadávanie	Vzor 1	Vzor 2	Vzor 3
Identické (ms)	2,526	2,174	2,204
Podštruktúrne (s)	49,5	15,425	14,433
Podštruktúrne* (s)	40,76	3,313	1,013

* Meranie zopakované po implementácii indexu pre odtlačky po dokončení diplomovej práce.



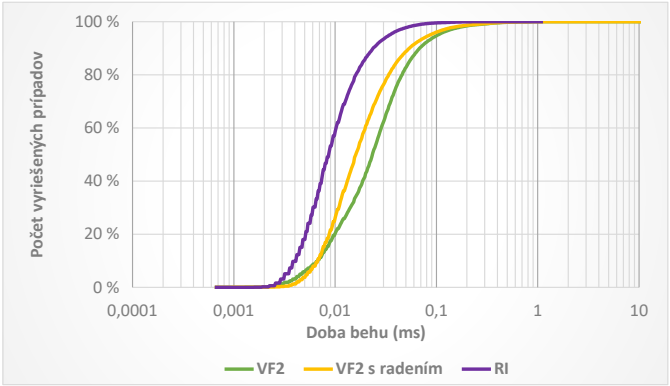
Architektúra systému

①



Porovnanie algoritmov

②



Algoritmus	Doba behu (µs)	Najrýchlejší (#)
VF2	36,284	720534
VF2 s radením	28,887	88661
RI	12,728	5476356
Kombinovaný*	11,976	

* Teoretický algoritmus, ktorý pri hľadaní použije pre daný prípad najrýchlejší implementovaný algoritmus. Ako je možné vidieť, zrýchlenie oproti RI je zanedbateľné.

Princíp odtlačku

Podštruktúrny odtlačok z knižnice RDKit je získaný vyhľadaním významných vzorov v štruktúre a ich následným hašovaním do bitového poľa. Výsledný odtlačok sa použije k rýchlemu vylúčeniu nevhodných záznamov. Presnosť odtlačku je približne 60 %, čo môže viesť k výraznému zredukovaniu počtu grafovo prehľadávaných štruktúr o viac ako 99 %.

③