Názov projektu:

Efektívne preferenčné Top-K vyhľadávanie (EpTopK)

Vedúci: Matúš Ondreička (matus.ondreicka <at> gmail.com)

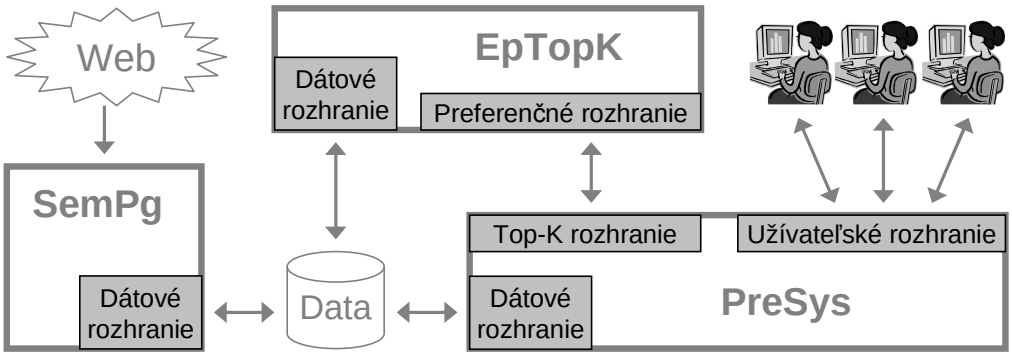
Jazyk, OS: doporučená je JAVA, ale môže byť aj iný

Počet riešiteľov: 4 až 5

Termín dokončenia: do 9 mesiacov od zahájenia projektu

Motivácia

V súvislosti s výskumom *Sémantického webu* a predovšetkým myšlienkou jeho postupnej *sémantizácie* [1] je potrebné vytvoriť rozsiahlu experimentálnu platformu W2U (Web to User). Súbežne s týmto softwarovým projektom sú preto vypísané dva nadväzujúce softwarové projekty, SemPg a PreSys, ktoré s budú zaoberať vývojom ďalších častí systému W2U. Extrakciou dát z webu a sa bude zaoberať softwarový projekt SemPg a problematikou užívateľských preferencií sa bude zaoberať softwarový projekt PreSys. Obrázok 1 znázorňuje platformu W2U a rozhrania medzi jeho jednotlivými časťami.



Obrázok 1

Popis projektu

Sémantický web sa zaoberá problematikou extrakcie dát z webu a následným spracovaním dát tak, aby užívatelia webu mohli v týchto dátach vyhľadávať podľa svojich užívateľských preferencií. Pritom väčšina užívateľov požaduje len malý počet najlepších výsledkov. Úlohou softwarového projektu EpTopK je vytvoriť systém, ktorý bude *v predspracovaných dátach efektívne vyhľadávať K najlepších objektov* (problém Top-K) *podľa užívateľských preferencií.*

Popis problému

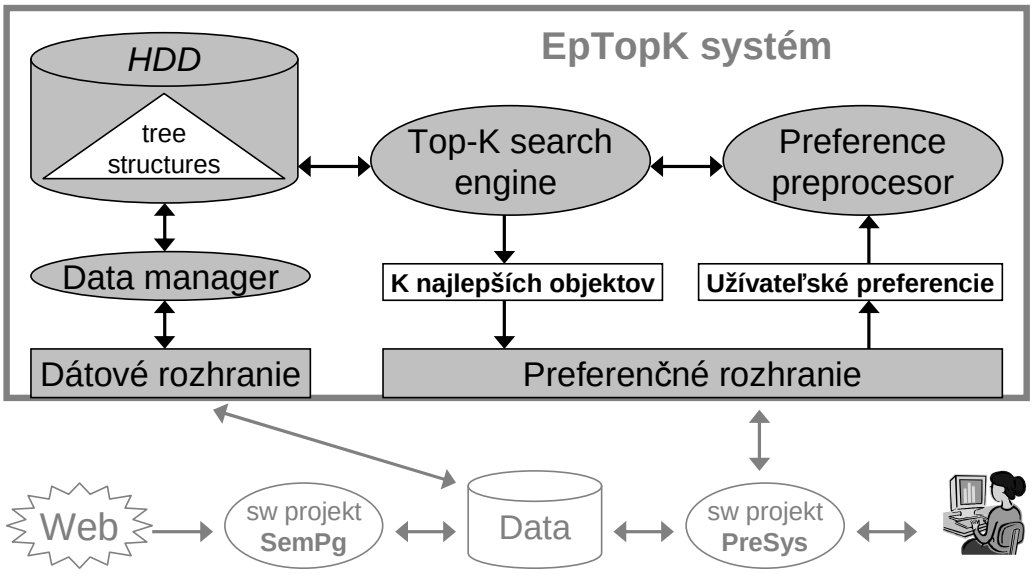
Predpokladá sa, že objekty sú rovnakého druhu (byty v Prahe) a majú viacero atribútov (napr. cena, rozloha, počet izieb, lokalita). Užívateľ sa potom podľa hodnôt týchto atribútov rozhoduje, ktoré objekty (byty) sú pre neho viac alebo menej vhodné. Každý užívateľ má pritom iné užívateľské preferencie, preferuje iné objekty.

Užívateľské preferencie sú modelované lokálne (fuzzy funkcie) a globálne (agregačná funkcie). Konkrétny užívateľ zadá ako preferuje objekty podľa jednotlivých atribútov (čím menšia cena, tým lepšie) a zároveň vyjadrí vzájomný vzťah atribútov (na cene záleží viac ako na lokalite). Úlohou je vytvoriť systém, ktorý bude efektívne vyhľadávať K (napr. 10) najlepších objektov pre užívateľa podľa jeho preferencií.

Aby bolo takéto vyhľadávanie pri danom modeli užívateľských preferencií efektívne, je potrebné použiť vhodné Top-K algoritmy [1], ktoré dokážu nájsť K najlepších objektov bez prehľadania všetkých objektov. Pre ich použitie je potrebné, aby boli dáta o objektoch vhodným spôsobom indexované v stromových dátových štruktúrach [2]. Tieto štruktúry musia byť vhodne uložené na pevnom disku.

Hrubý návrh riešenia

Na obrázku 2 je načrtnutý jeden z možných návrhov architektúry systému EpTopK. V systéme by mali byť implementované komponenty, ktoré sú na obrázku vyfarbené šedou farbou. Šípky znázorňujú tok dát v systéme.



Obrázok 2

Popis komponent

- **Dátové rozhranie** – umožní systému získavať dát o objektoch z externých zdrojov
  - Ďalej by mal systém dokázať získavať dáta z ďalších zdrojov (napr. komunikácia vo formáte XML).

- **Data manager** –spracovanie a obsluha získaných dát
  - Ukladá získané dáta do interných stromových dátových štruktúr systému, na pevný disk.
  - V rozumnej miere by mal dokázať aktualizovať svoje dáta vzhľadom ku externým dátam (napr. raz denne).
  - Mal by dynamicky optimalizovať stromové dátové štruktúry vzhľadom ku konkrétnym dátam, ktoré sa v nich nachádzajú (napr. distribúcia dát).

- **HDD** – dátové úložisko systému (na pevnom disku)
  - Úlohou riešiteľov projektu EpTopK bude vhodne navrhnúť a implementovať stromové dátové štruktúry na pevnom disku tak, aby bol počas výpočtu algoritmov bol počet prístupov na pevný disk čo najmenší.

- **Preprocesor preferencií** – spracovanie preferencií pre Top-K vyhľadávanie
  - Bude obsahovať všetky potrebné metódy pre podporu Top-K vyhľadávania.

- **Top-K search engine** – komponenta Top-K vyhľadávania
  - Bude obsahovať algoritmy pre Top-K vyhľadávanie, ktoré budú vyhľadávať v dátovom úložisku systému podľa preferencií.
  - Bude obsahovať metódy na získavanie objektov z dátového úložiska.
  - Úlohou riešiteľov projektu EpTopK bude vhodne navrhnúť, ako sa budú pripravovať vstupy pre implementované Top-K algoritmy (napr. nová komponenta).
  - Pre zvyšovanie efektívnosti vyhľadávania by mal systém vhodným spôsobom merať výsledky Top-K algoritmov, objem získaných dát z dátového úložiska systému, čas výpočtu algoritmov a pod..

- **Preferenčné rozhranie** – získavanie preferencií
  - Ďalšou časťou systému bude implementácia preferenčného rozhrania, ktoré umožní serveru EpTopK dodať užívateľské preferencie (vstup) a umožniť mu odoslať K najlepších objektov (výstup). Pomocou tohto rozhrania by mal byť server EpTopK prepojiteľný s inými projektmi.

Vzťah ku ostatným softwarovým projektom

Počas implementácie tohto softwarového projektu pravdepodobne ešte nebudú plne implementované ostatné softwarové projekty. V tom prípade bude potrebné vytvoriť triviálne komponenty, ktoré bude možné pripojiť na rozhrania systému EpTopK.

Napríklad pre ďalšie praktické využitie systému EpTopK je vhodné, aby systém získaval dáta z relačnej databázy (dátové rozhranie).

Rovnako pre preferenčné rozhranie by bolo pre ďalšie praktické využitie systému EpTopK vytvoriť jednoduché grafické užívateľské prostredie, v ktorom budú užívatelia zadávať svoje preferencie a v ktorom sa im budú zobrazovať požadované najlepšie objekty (napr. formou web stránky).

Poznámka

Tento softwarový projekt je vhodný pre študentov, ktorý by chceli na projekte (na ďalšom návrhu/vývoji jeho častí) ďalej pokračovať v diplomovej práci.

Referencie

[1] Jan Dědek, Alan Eckhardt, Leo Galamboš, Peter Vojtáš: *Sémantický Web*, in DATAKON 2008, Brno, ISBN 978-80-7355-081-3, pp. 12-30, 2008.

[2] Ondreička, M., Pokorný J.: *Extending Fagin's algorithm for more users based on multidimensional B-tree*. In: Proc. of ADBIS 2008, P. Atzeni, A. Caplinskas, and H. Jaakkola (Eds.), LNCS 5207, Springer-Verlag Berlin Heidelberg, 2008, pp. 199–214.