Personalizované vyhľadávanie v zdrojovom kóde

Eduard Kuric

prof. Mária Bieliková

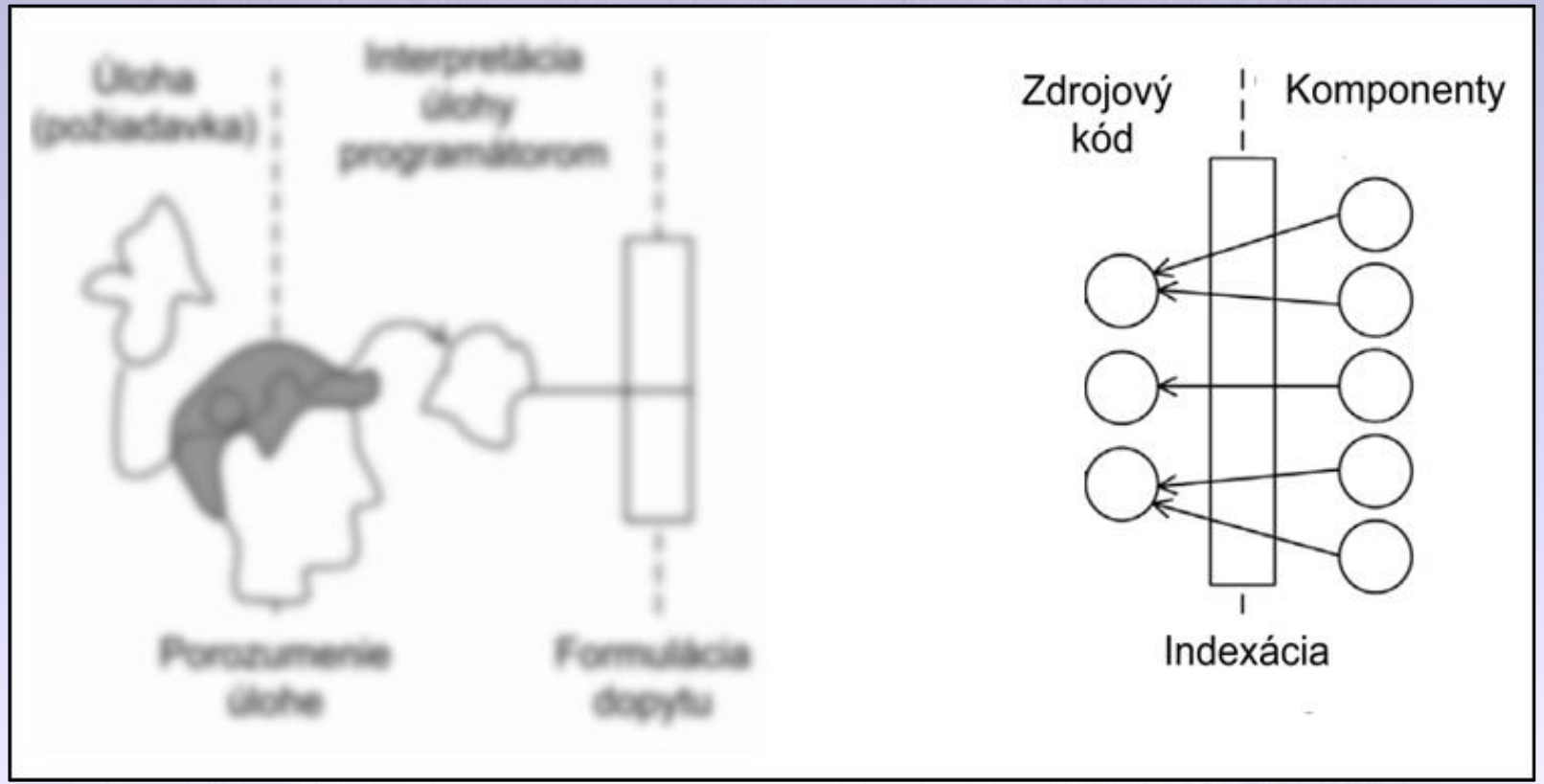
Motivácia

- **Web**
  - text (stránky), obrázky, videá, hudba -> prirodzený jazyk
- **web ako pavučina softvérových artefaktov $\subset$ Web**
  - repozitár zdrojových kódov
- **vyhľadávanie v zdrojových kódoch**
  - znovupoužitie, vykonanie zmien
  - (potenciálne) riešenie, inšpirácia, oživenie spôsobu použitia

→ **nepostačuje jednoduchý "plnotextový" vyhľadávač**

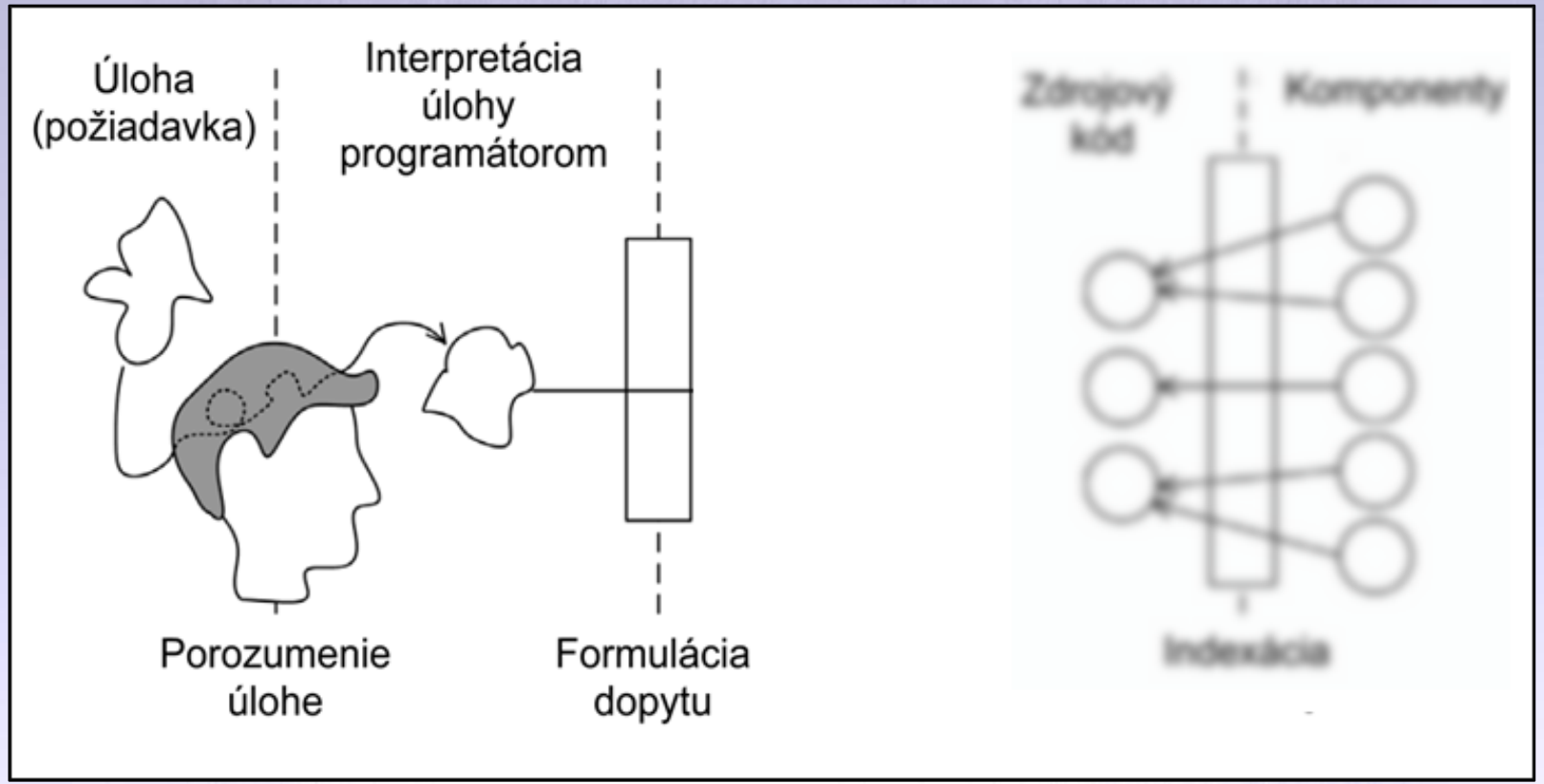
  - kognitívne bariéry, sociálny aspekt

Model vyhľadávania komponentov



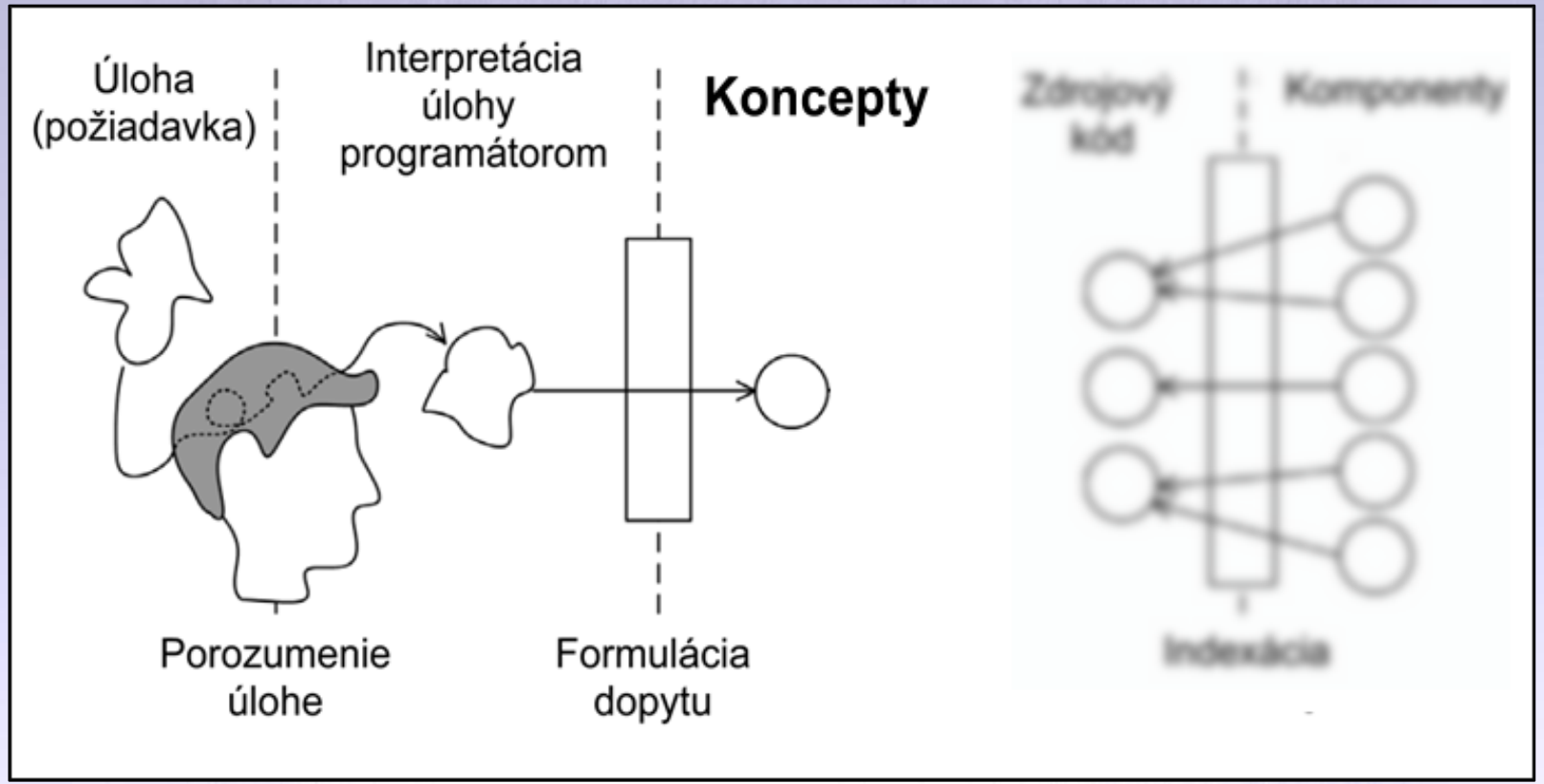
(Lucredio, et al., 2004)

Model vyhľadávania komponentov



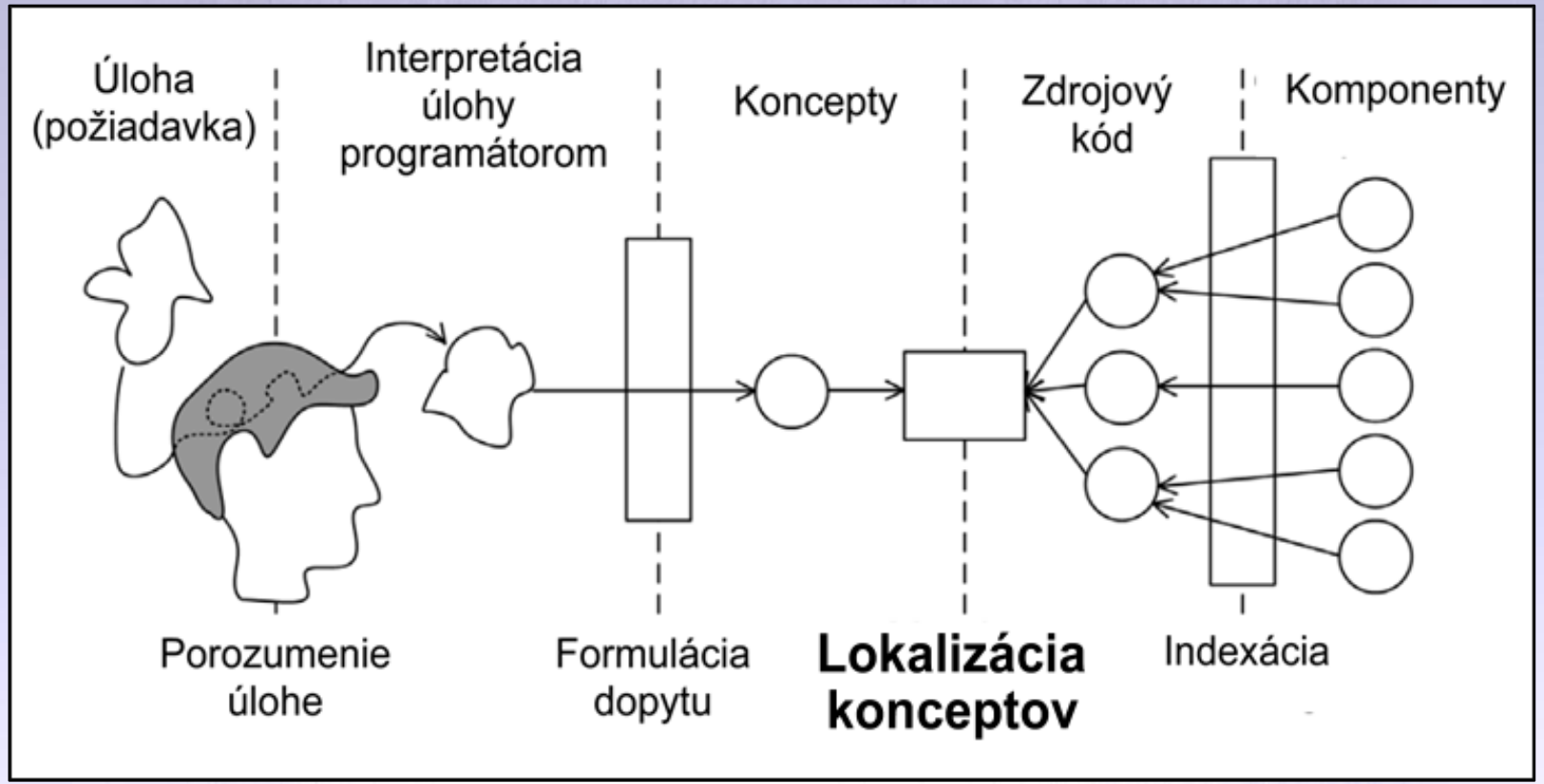
(Lucredio, et al., 2004)

Model vyhľadávania komponentov



(Lucredio, et al., 2004)

Model vyhľadávania komponentov



(Lucredio, et al., 2004)

Existujúce prístupy – lokalizácia konceptov

- **Dynamické prístupy**
  - vyžadujú spustiteľné programy
  - sledovanie a analyzovanie vykonávania programu
  - identifikácia vlastností systému
    (Bogdan et al., survey, 2011)

- **Statické prístupy**
  - zdrojový kód je spracovaný ako „textový“ dokument
  - založené na známych metódach z oblasti vyhľadávania informácií, napr. porovnávanie vzorov, VSM, LSI+FCA, LDA
    (Zhao et al., 2006; Poshyvanyk et al., 2007; Bajracharya & Lopes, 2009)

Úskalia existujúcich prístupov

- Existujúce prístupy sa zameriavajú spravidla na presnosť:
  - lokalizovanie komponentov, ktoré obsahujú čo najväčší počet konceptov z dopytu;
  - na redukciu priestoru, ktorý potrebuje programátor preskúmať.

→ **Kognitívne bariéry** (Ye et al., 2002; Lucredio, et al., 2004; Gay et al., 2009)
  - nedostatok znalostí o softvérovom systéme
  - nepoužívanie potenciálne užitočných komponentov
  - konceptuálne medzery (drawCircle vs. drawOval)

Reputácia zdrojového kódu

- web – rozsiahly repozitár zdrojových kódov
- Jednoduchý „plnotextový“ vyhľadávač nepostačuje
  - popri relevancii je potrebné zohľadňovať **reputáciu zdrojových kódov**

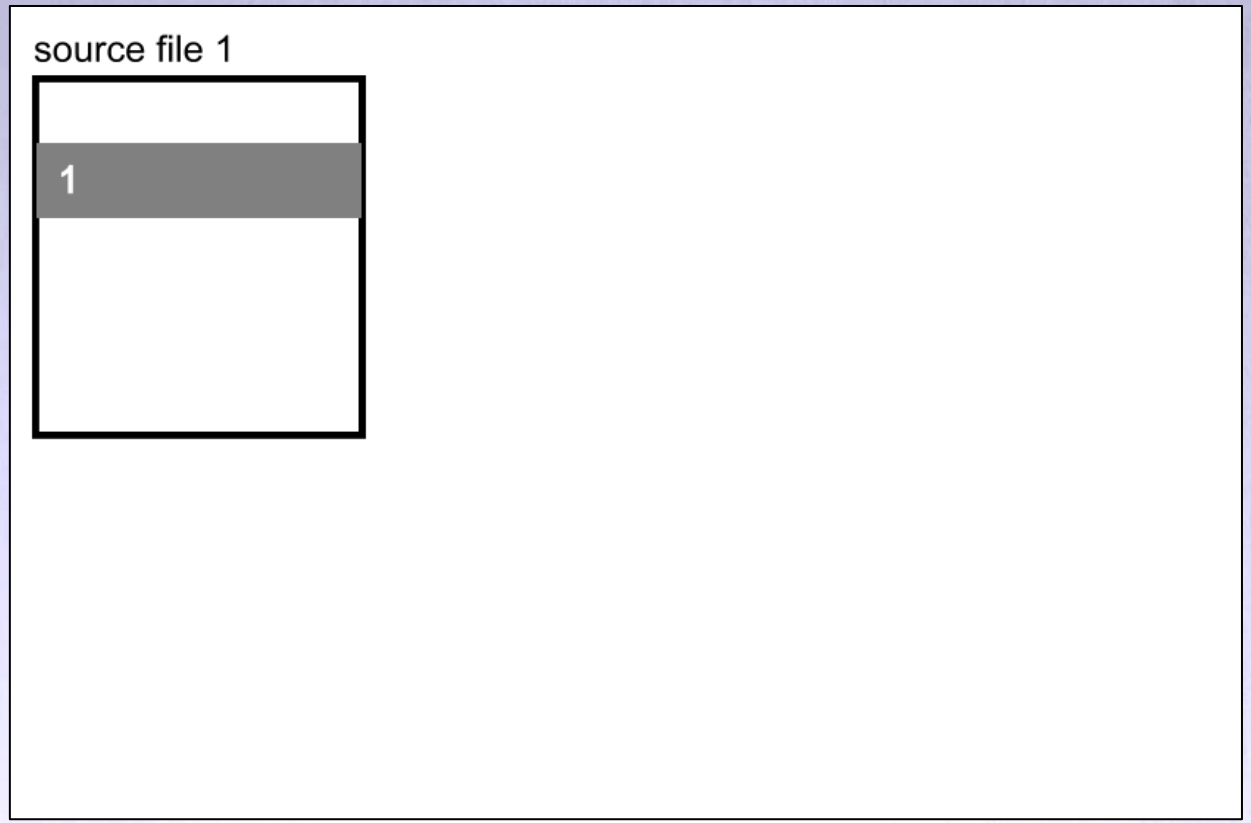
- Stanoviská (názory, skúsenosti) iných programátorov
- Štatistiky o projekte

- **Reputácia autorov zdrojových kódov?**
  → model skúseností programátora z jeho aktivít pre personalizované vyhľadávanie

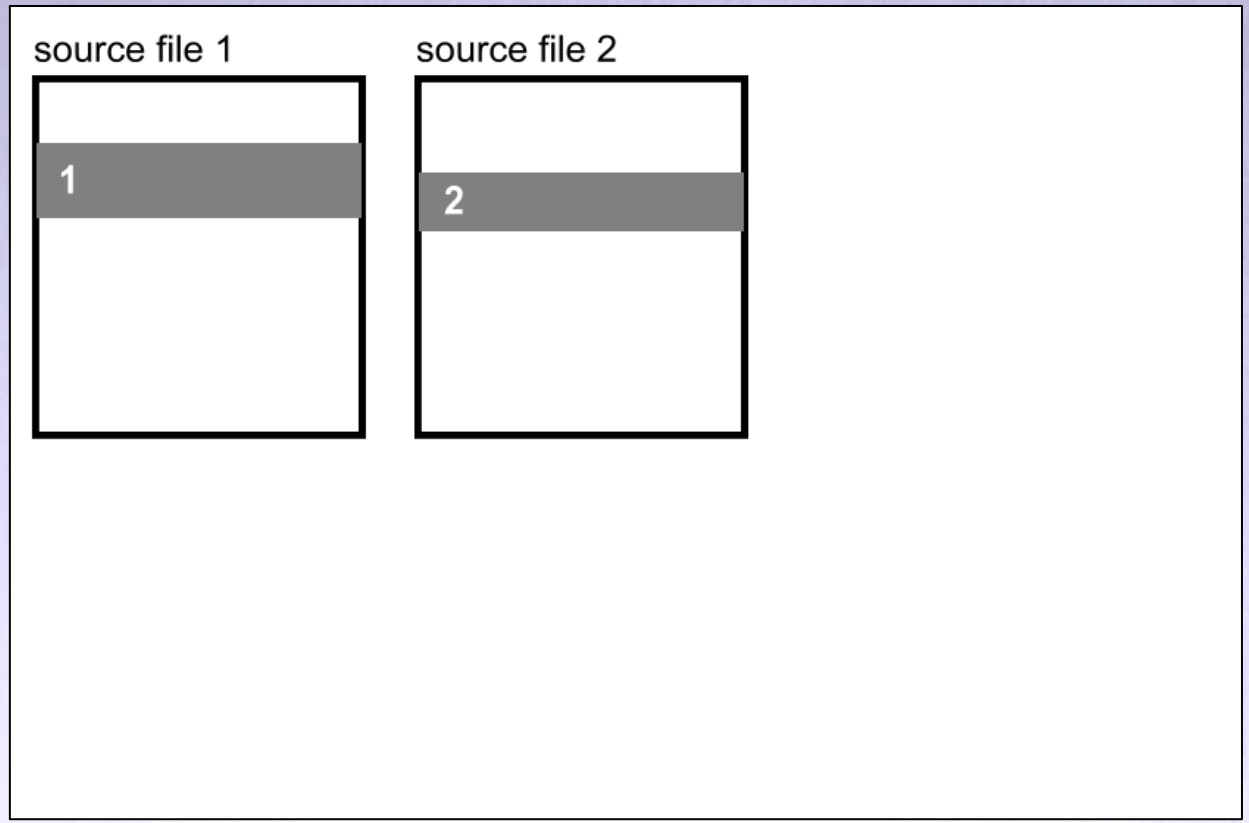
Modelovanie skúseností programátora

- Autorstvo
- Nízkoúrovňové aktivity (IDE)
- Význam komponentu
- Stabilita komponentu (trvanie autorstva)
- Skúsenosti programátora (technológie)

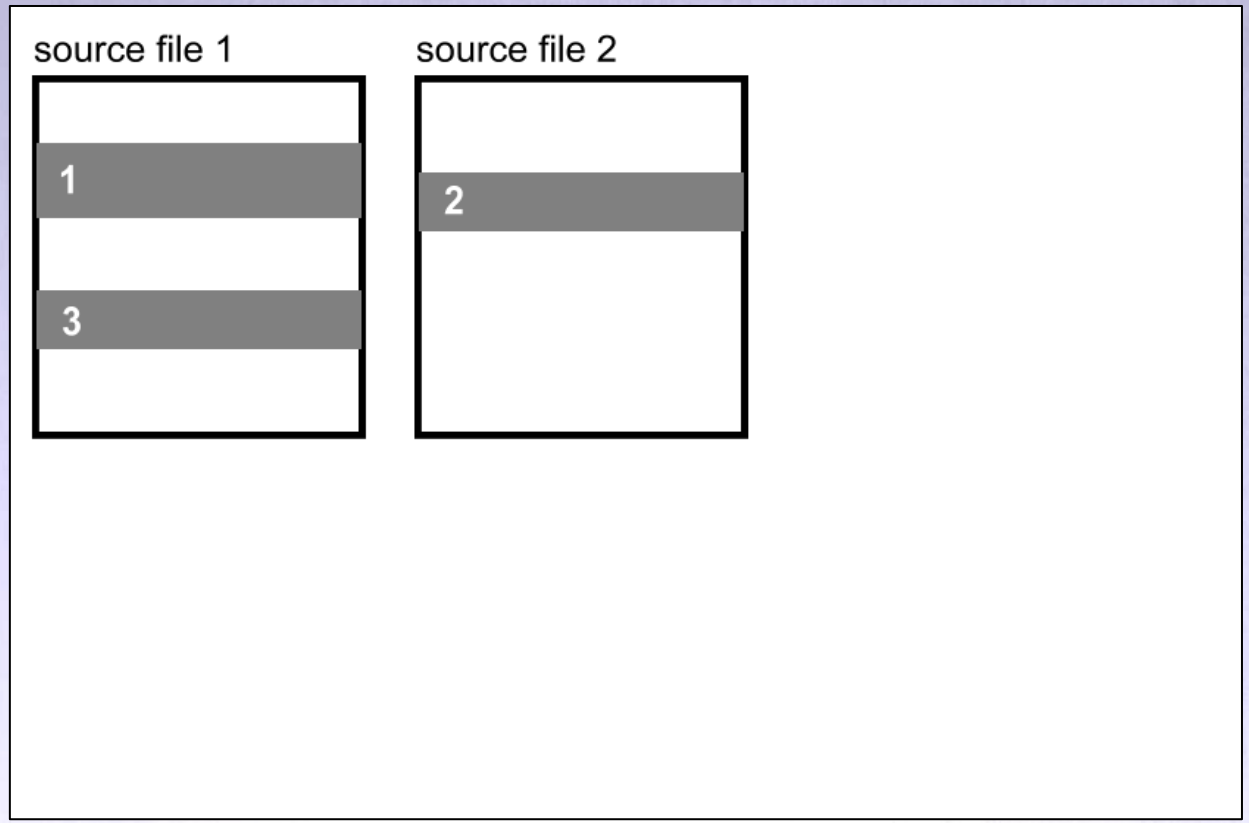
Scenár vyhľadávania



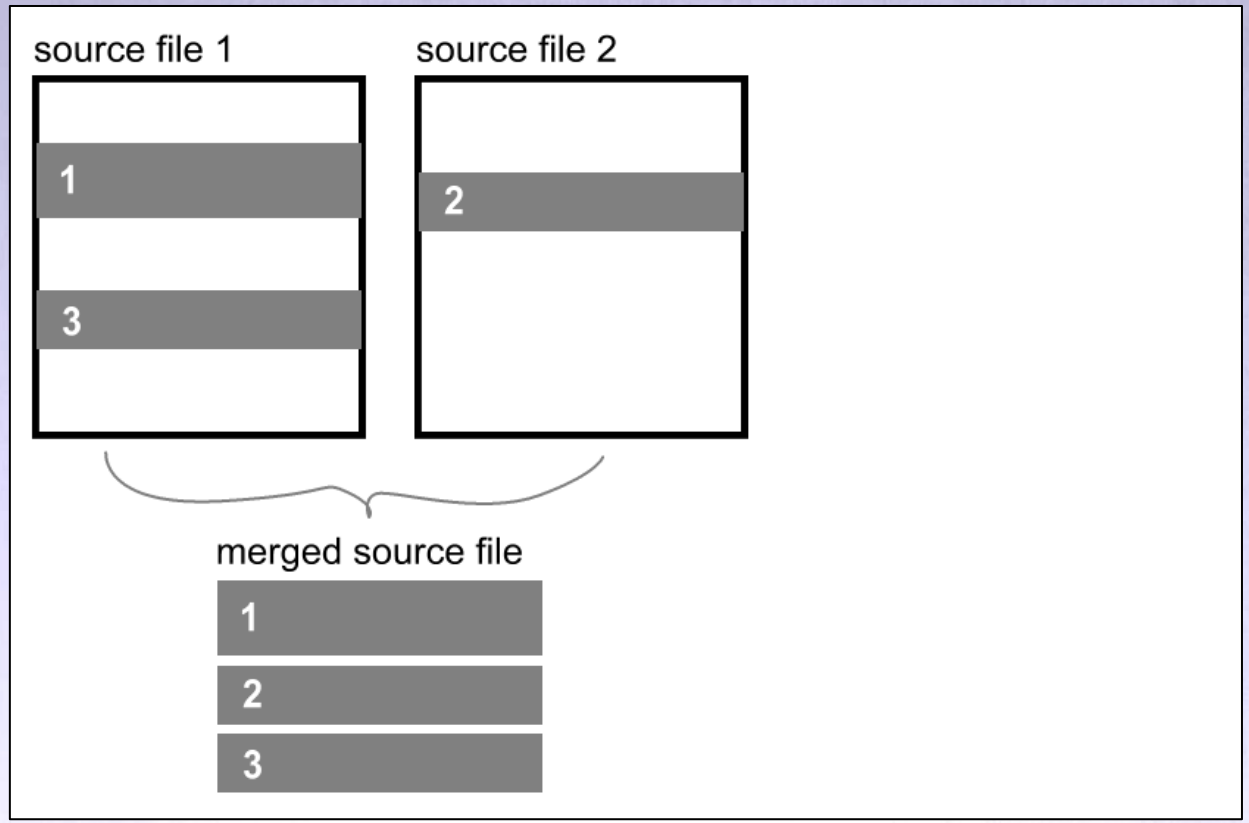
Scenár vyhľadávania



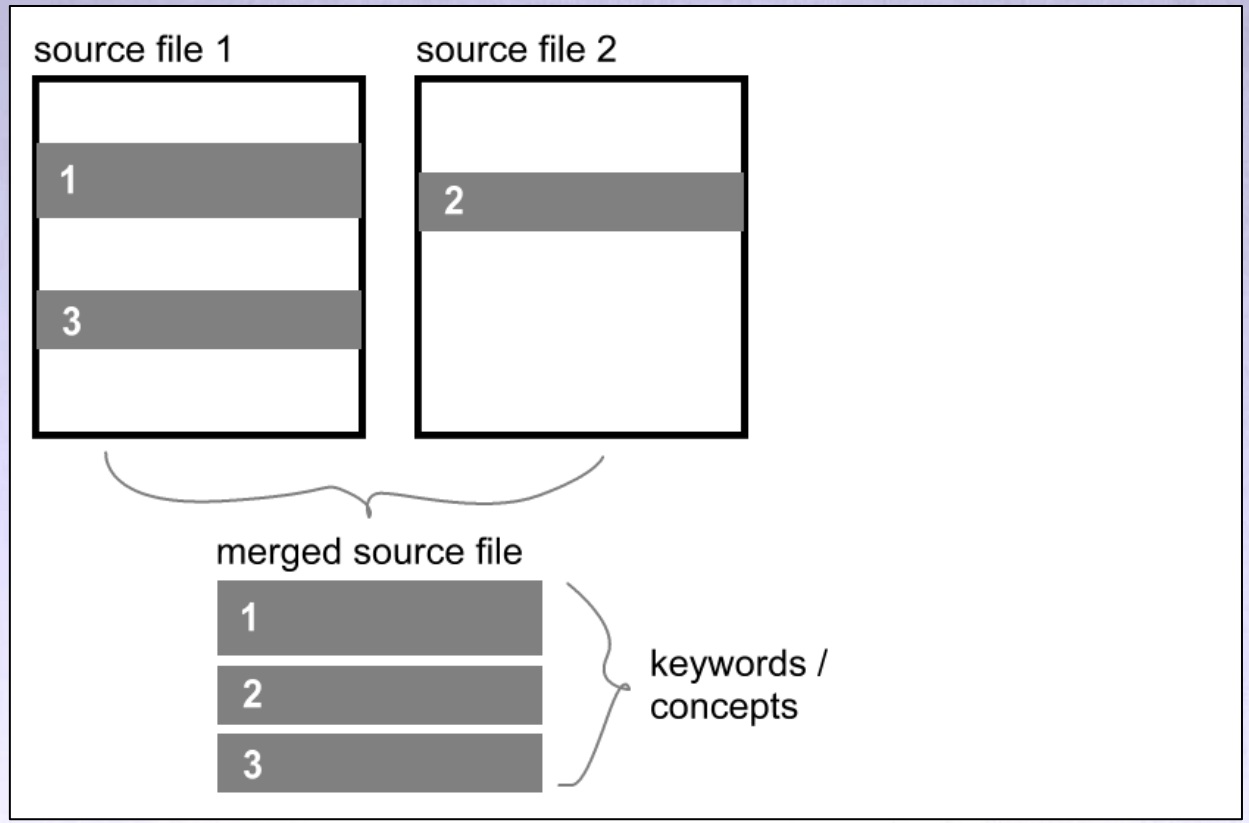
Scenár vyhľadávania



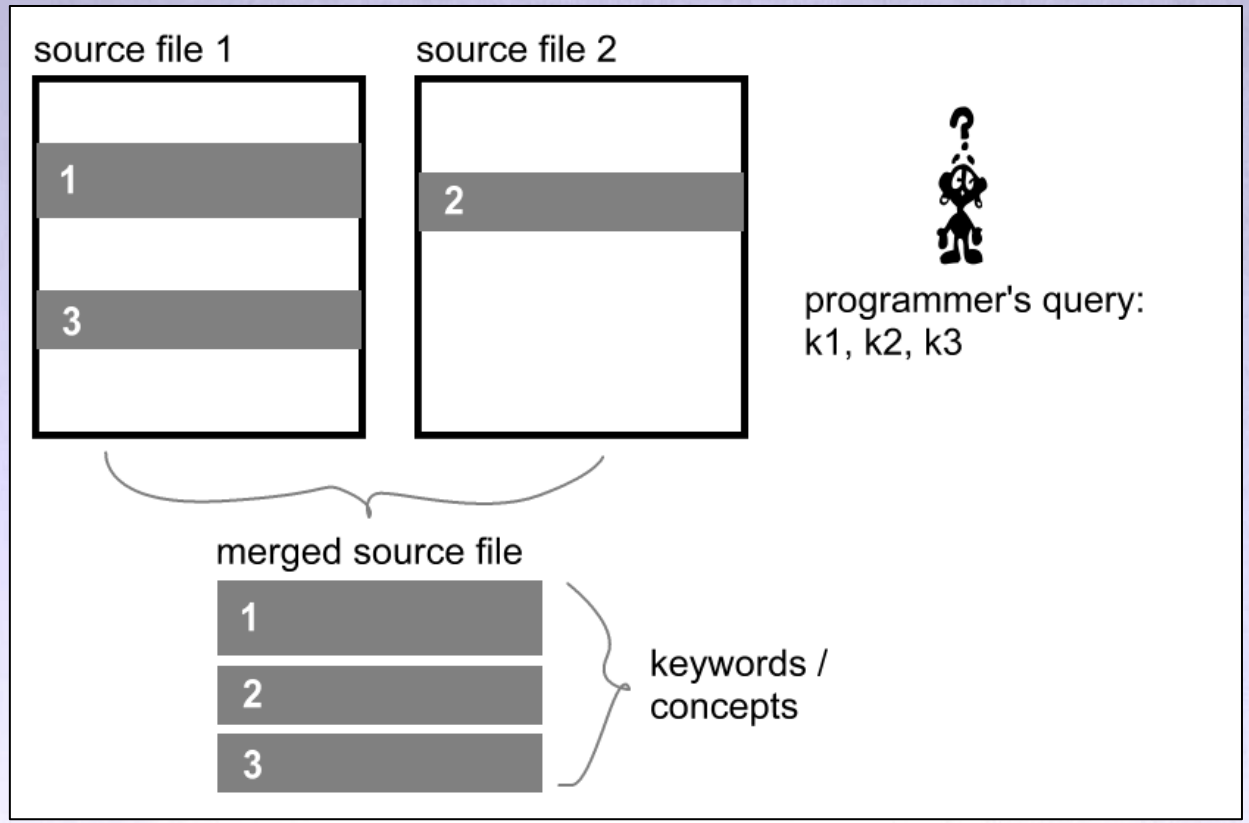
Scenár vyhľadávania



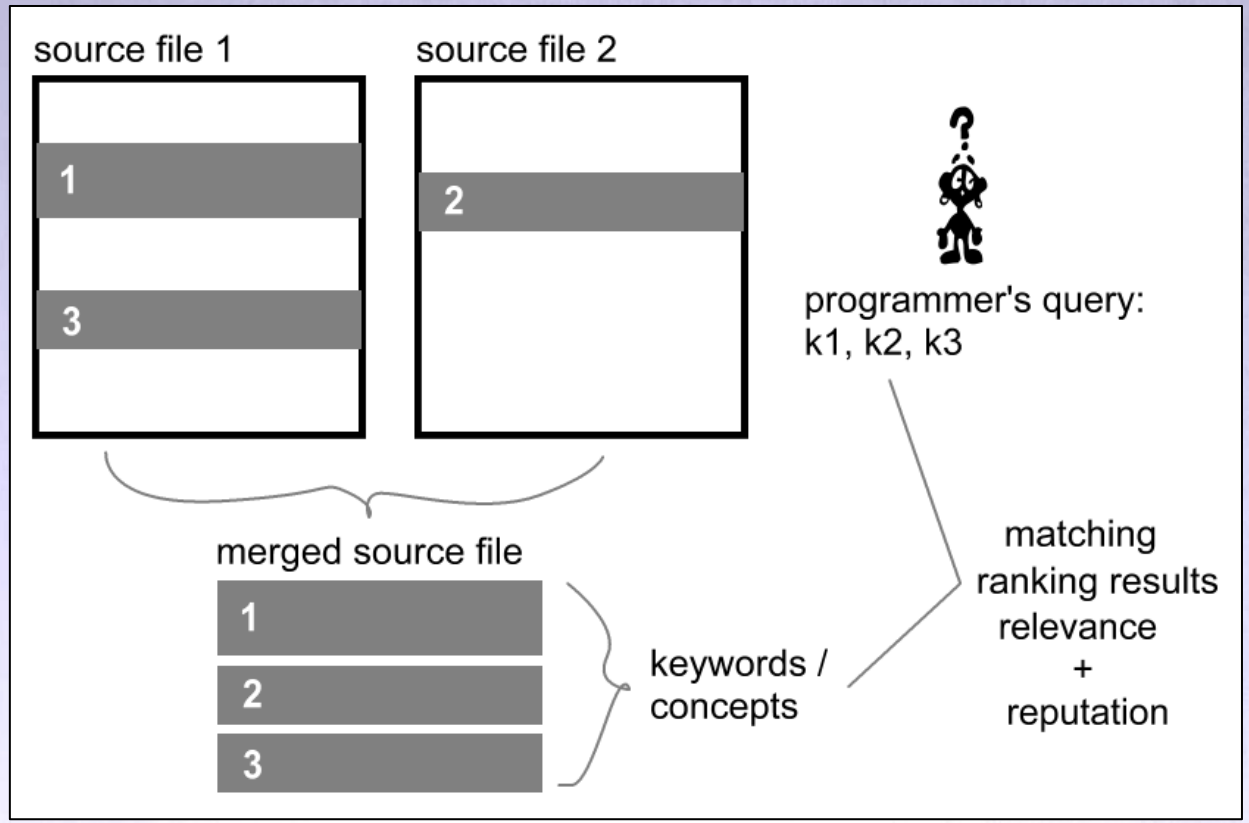
Scenár vyhľadávania



Scenár vyhľadávania



Scenár vyhľadávania



Aktuálny stav – navrhnuté metódy

- **Vyhľadávanie v zdrojovom kóde na základe identifikácie populárnych fragmentov**
  - Identifikácia populárnych komponentov v zdrojových kódoch
  - PageRank algoritmus
  - Popularita komponentu (metódy) je určená na základe počtu komponentov odkazujúcich na cieľový komponent

- Prijatý článok: SOFSEM 2013, Springer LNCS

Aktuálny stav – navrhnuté metódy /2

- **Automatické extrahovanie skúseností programátorov zo zdrojových kódov**
  - Ktoré knižnice (technológie) programátori reálne používajú, kedy ich naposledy použili a do akej miery (hĺbky)
  - Identifikácia technológií zo zdrojových kódov (napr. Hibernate)
  - Odhad skúseností programátora s technológiou v porovnaní s ostatnými programátormi

- Prijatý článok: ZNALOSTI 2012