Hľadanie vzťahov medzi kľúčovými slovami

Peter Kajan

Vedúci práce: Ing. Michal Barla, PhD

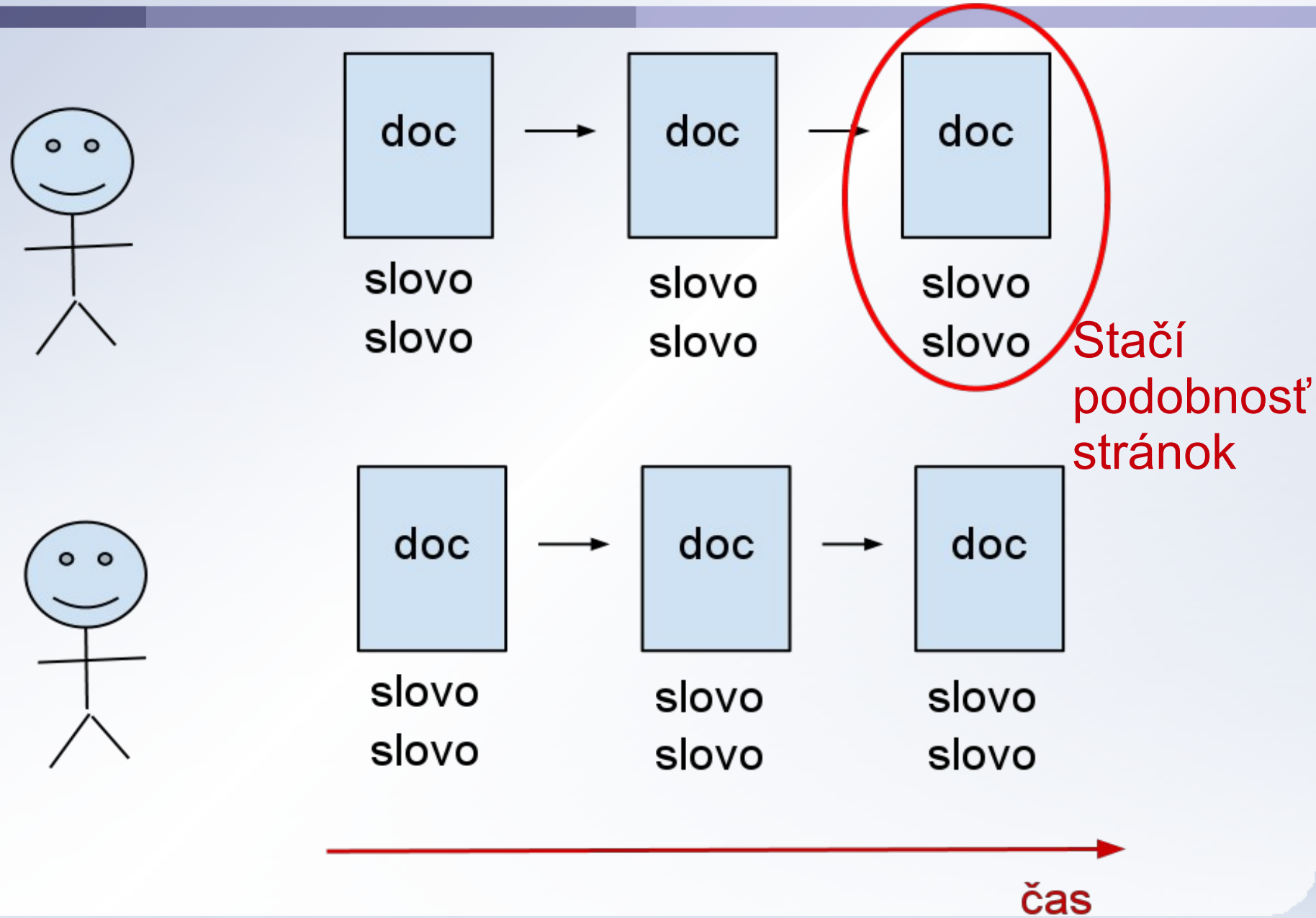
Fakulta informatiky a informačných technológii
Slovenská technická univerzita

Motivácia

- Cieľ – vzťahy slov na základe používania dokumentov

- Príbuznosť objektov
- Klasifikácia
- Odporúčanie tagov
- Znalostná báza

Charakter dát



Charakter dát (2)

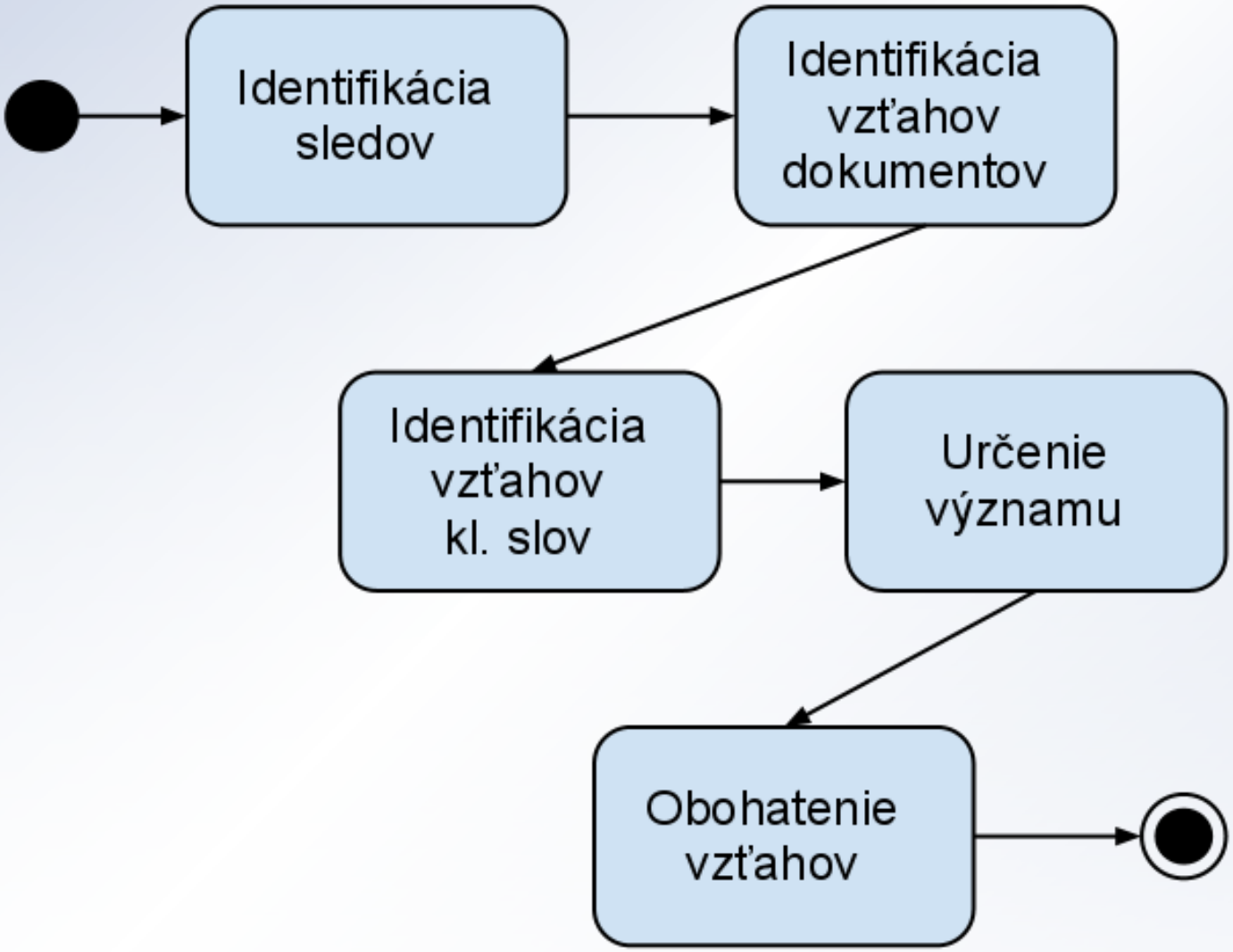
Výhody:

- Lacnejšie
- Pokrývajú viac domén

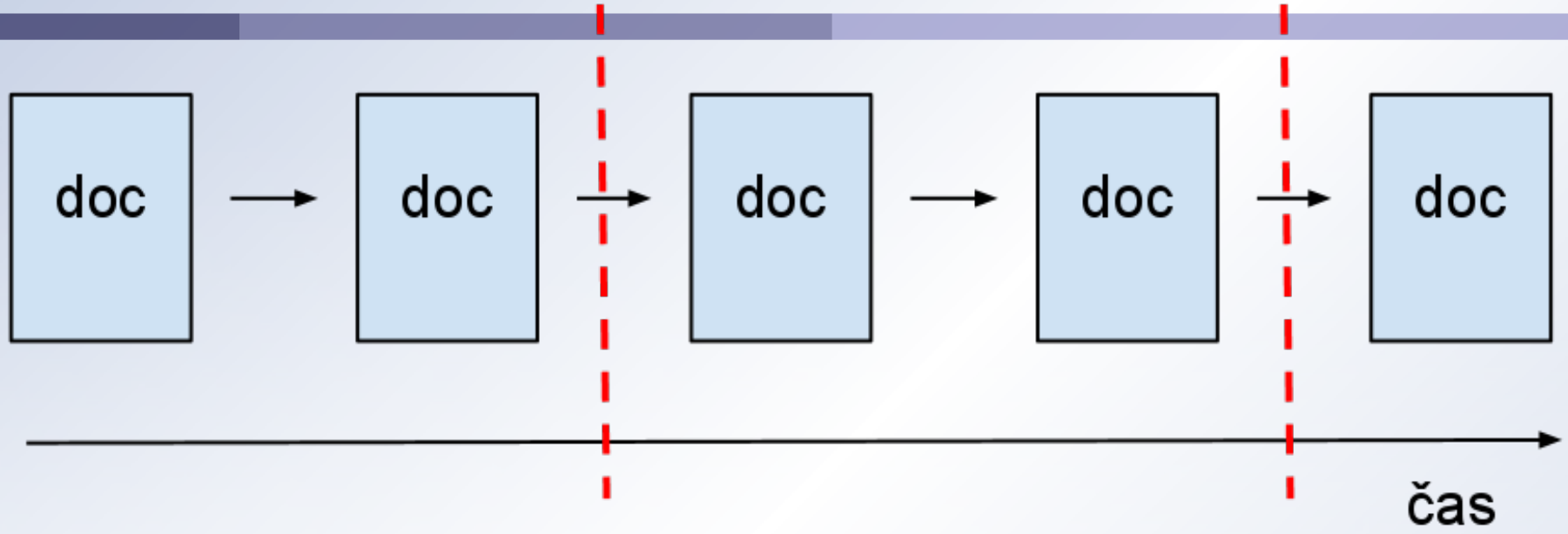
Nevýhody:

- Menšia rôznorodosť – tvoria tvorcovia, používatelia prezerajú

Návrh metódy



Identifikácia sledov



- Koincidencia vrámci dokumentov
- Časové okno

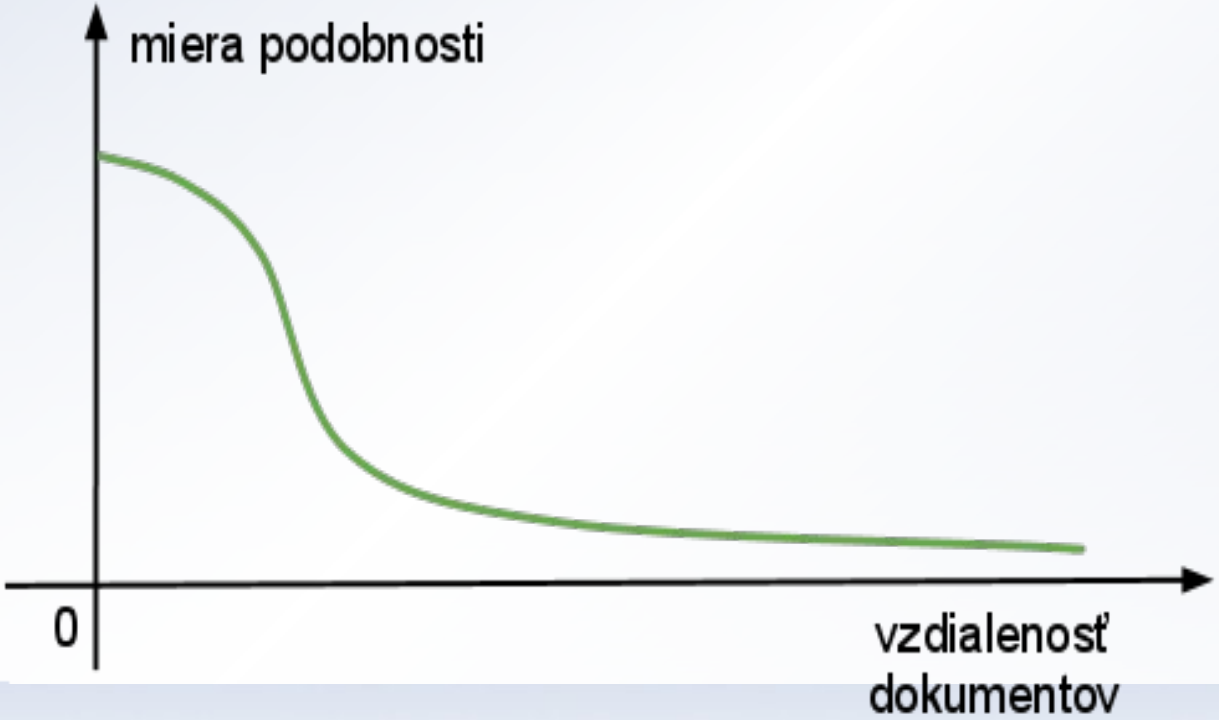
Zvažované:

- Existujúca znalostná báza (ConceptNet, Linked data)
- Naša znalostná báza – iteratívne vylepšovanie

Podobnosť stránok

- Hypotéza: čím bližšie pri sebe, tým viac spolu súvisia

$$s_d(d_a, d_b) = \sum^{\forall (a,b) \in Q_{ab}} f(|a-b|)$$



Koncept príslušnosti

- ka - výskyt *x*
- kb – výskyt y

x↑ c(ka,kb)↓
y↓ c(kb,ka)↑

- Príslušnosť *auta* k *Yaris*
- Príslušnosť *Yaris* k *autu*
- Rovnako pre dokumenty

Koncept príslušnosti (2)

- Príslušnosť dokumentov

$$c_d(d_a, d_b) = idf(d_a) \cdot s_d(d_a, d_b)$$

- Príslušnosť kľúčových slov

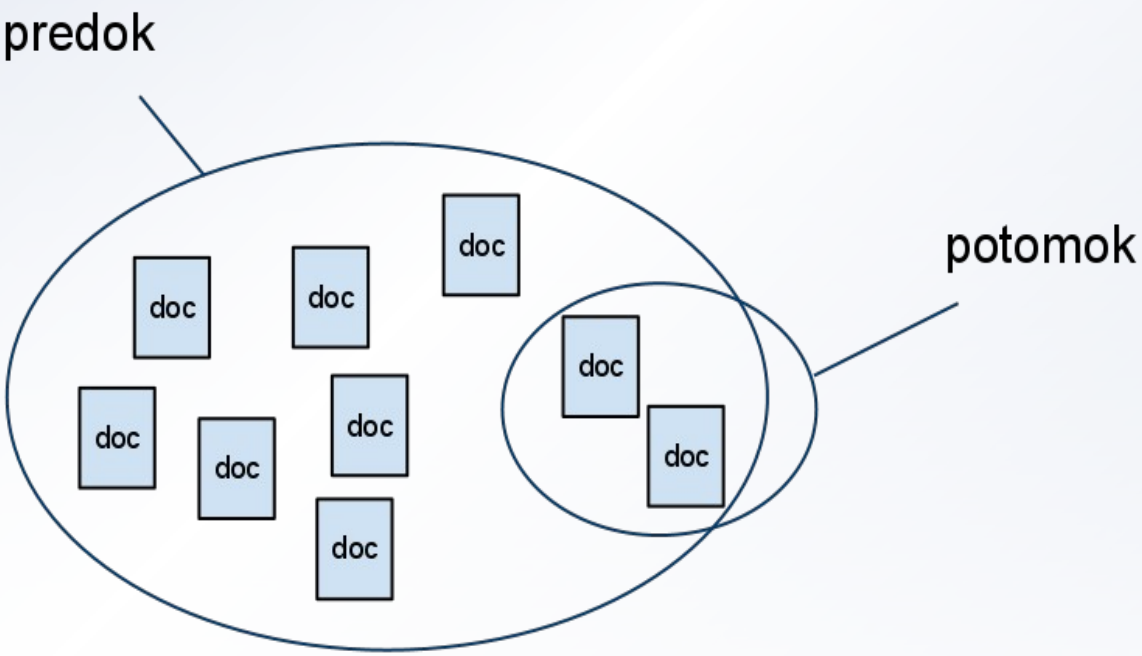
$$c_k(k_a, k_b) = k \cdot idf(k_a) \cdot \sum^{\forall d_a \in A, \forall d_b \in B} c_d(d_a, d_b)$$

- Dôsledky:
  - Penalizácia častých slov a slov nachádzajúcich sa v častých dokumentoch
  - Rôznorodejší priestor využiteľný pri tvorbe hierarchie
- Podobnosť kl. slov – geometrický priemer príslušností

Určenie hierarchie

- Pravdepodobnosť hierarchie z príslušnosti

$$p_h(k_b, k_a) = m \cdot \frac{c_k(k_a, k_b)}{c_k(k_b, k_a)}$$



Odstránenie problému menej používaného synonyma

- Hypotéza: Potomok slova bude v sledoch skôr za predkom
- Priemerná inverzná vzdialenosť

$$aid(k_a, k_b) = \frac{\sum^{\forall (a,b) \in R} \frac{1}{a-b}}{|R|}$$

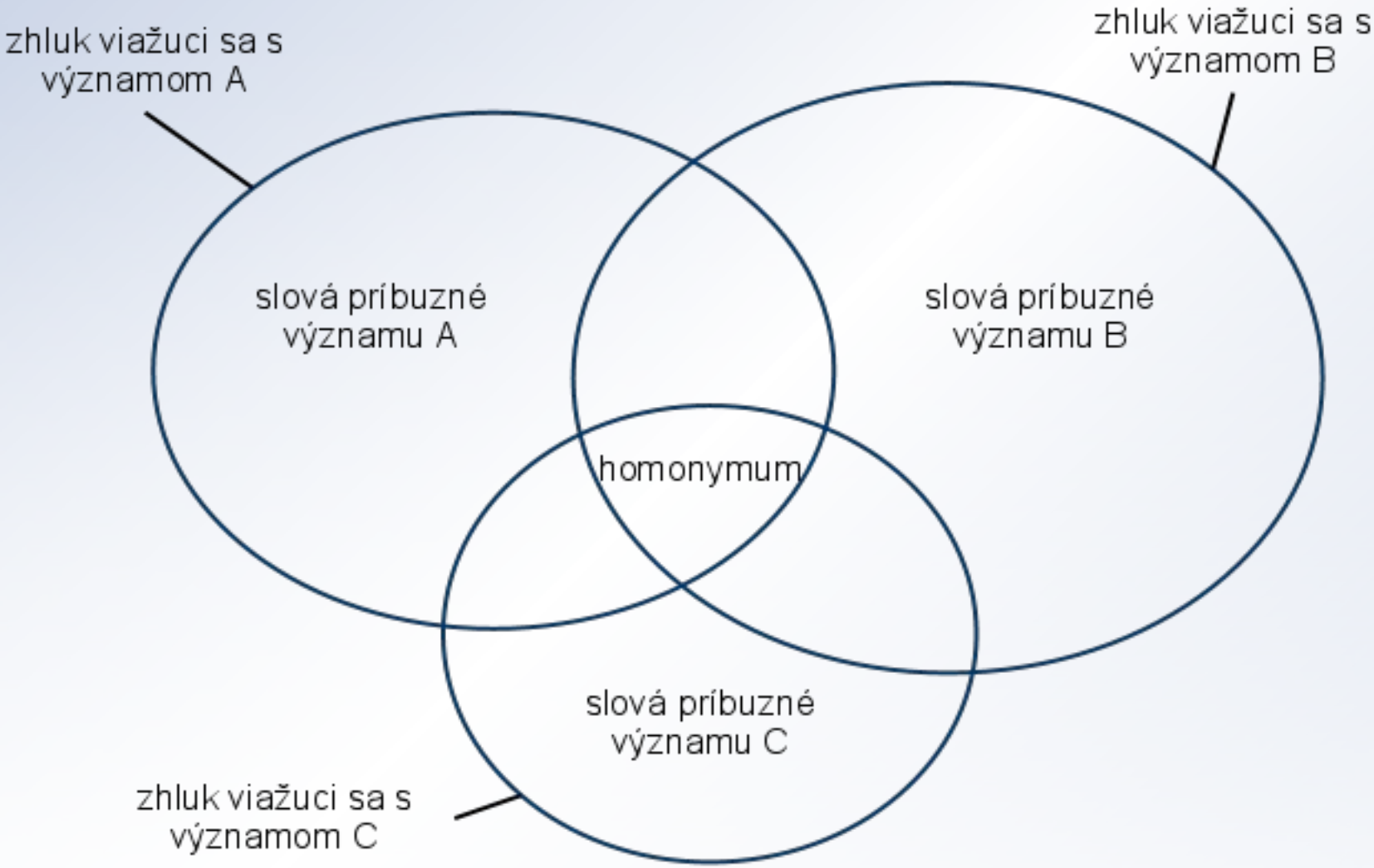
- < 0 *ka* sa nachádza pred *kb*
- > 0 *ka* sa nachádza za *kb*

Pravdepodobnosť hierarchie

$$p_h(k_b, k_a) = m \cdot \frac{c_k(k_a, k_b)}{c_k(k_b, k_a)} + n \cdot aid(k_b, k_a)$$

- Ak prekročí stanovený prah, určí sa vzťah

Určenie významu – získanie kontextu



Využitie Linked Data

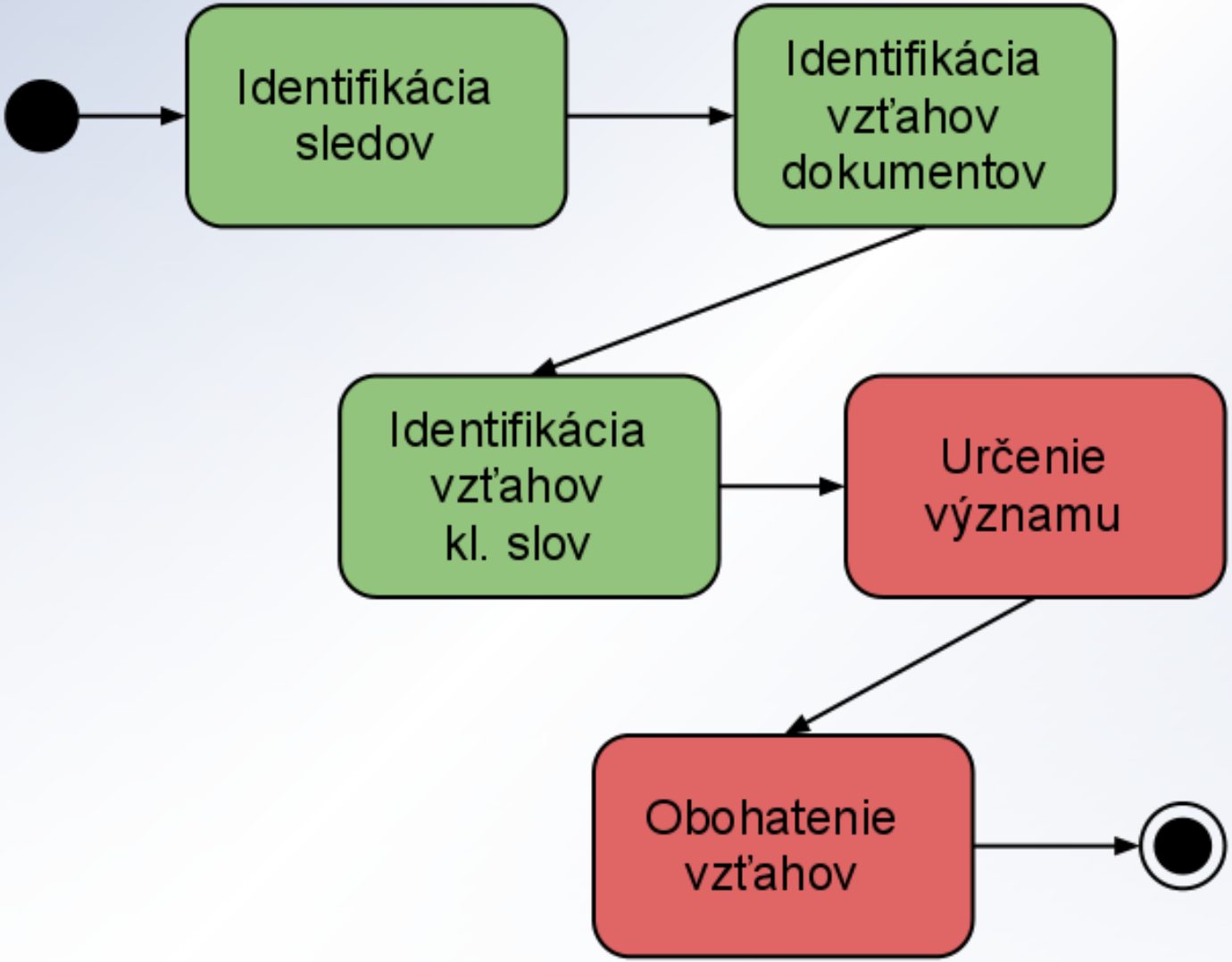
- Namapovanie na Linked Data na základe kontextu
- Využitie služieb (Open Calais) alebo iných existujúcich riešení
- Získanie vzťahov z linked data

```
SELECT distinct(?p)
WHERE {<UriA> ?p <UriB>}
```

Overenie riešenia

- Kvalitatívne, formou dotazníkov
- Zamerané na:
  - overenie podobností kľúčových slov
  - overenie presnosti a pokrytia vzťahov
  - overenie vplyv časovej zložky

Stav projektu



Zhrnutie

- Návrh metódy na automatické vytvárania znalostnej bázy
- Vzťahy kľúčových slov z používania dokumentov
- Využitie a overenie prínosu časovej informácie
- Riešenie problému identifikácie hierarchie
- Overenie v doméne používania webových stránok