Hľadanie vzťahov medzi kľúčovými slovami

Peter Kajan

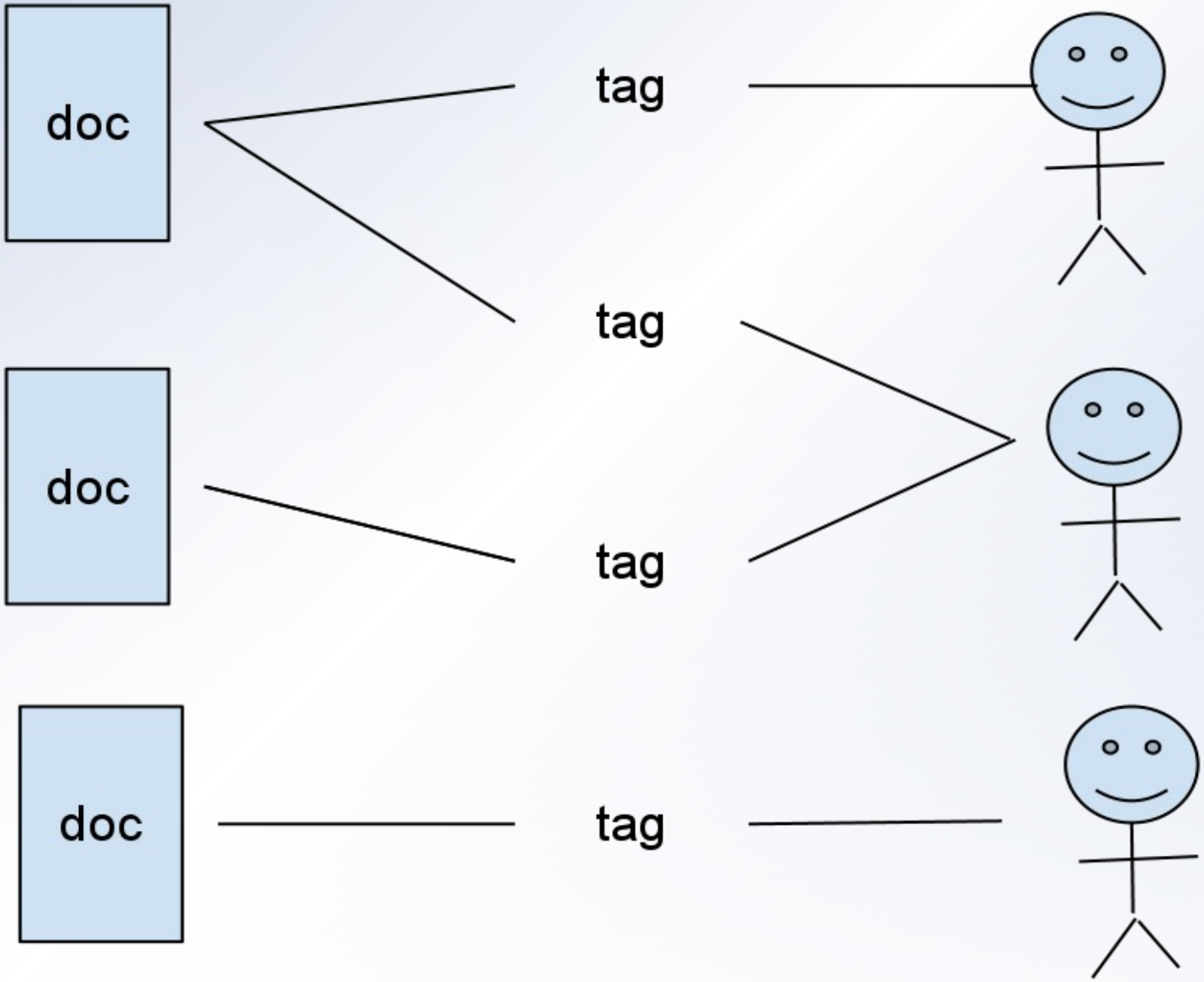
Vedúci práce: Ing. Michal Barla, PhD

Fakulta informatiky a informačných technológii
Slovenská technická univerzita

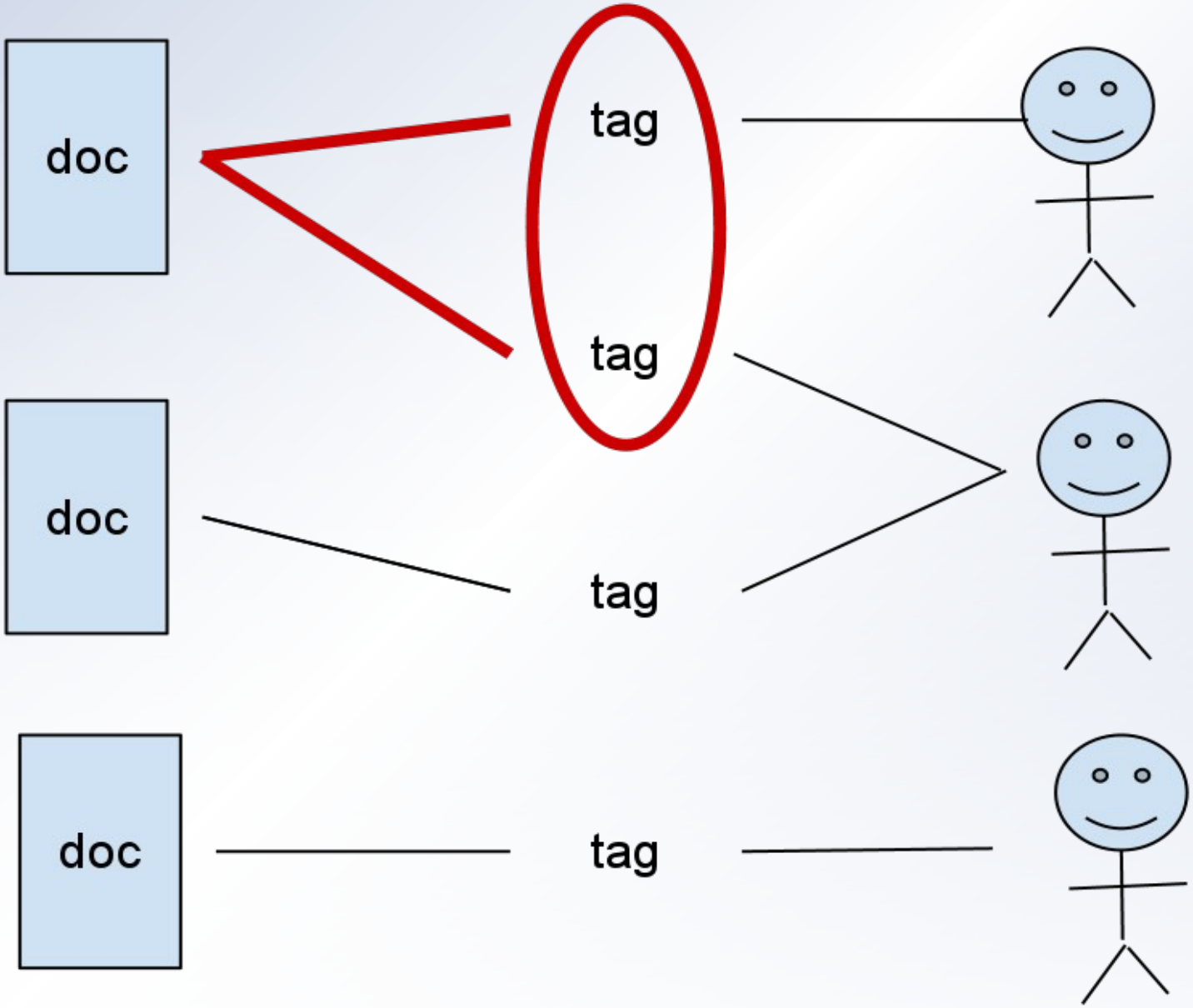
Motivácia

- Príbuznosť objektov
  - Dokumentov (stránok)
  - Používateľov
  - Komunít …
- Klasifikácia
  - …
- Odporúčanie tagov
- Znalostná báza

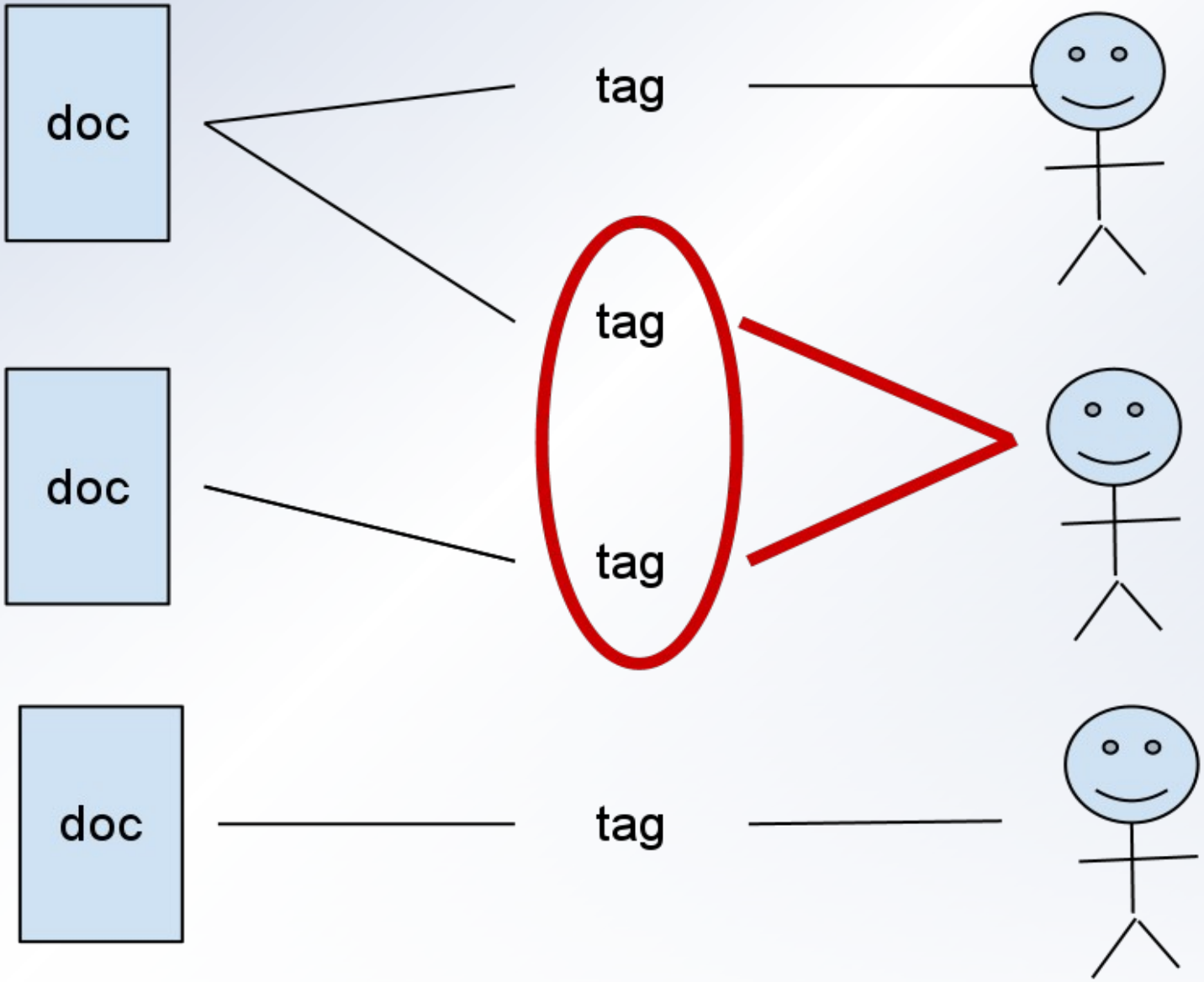
Predošlé prístupy



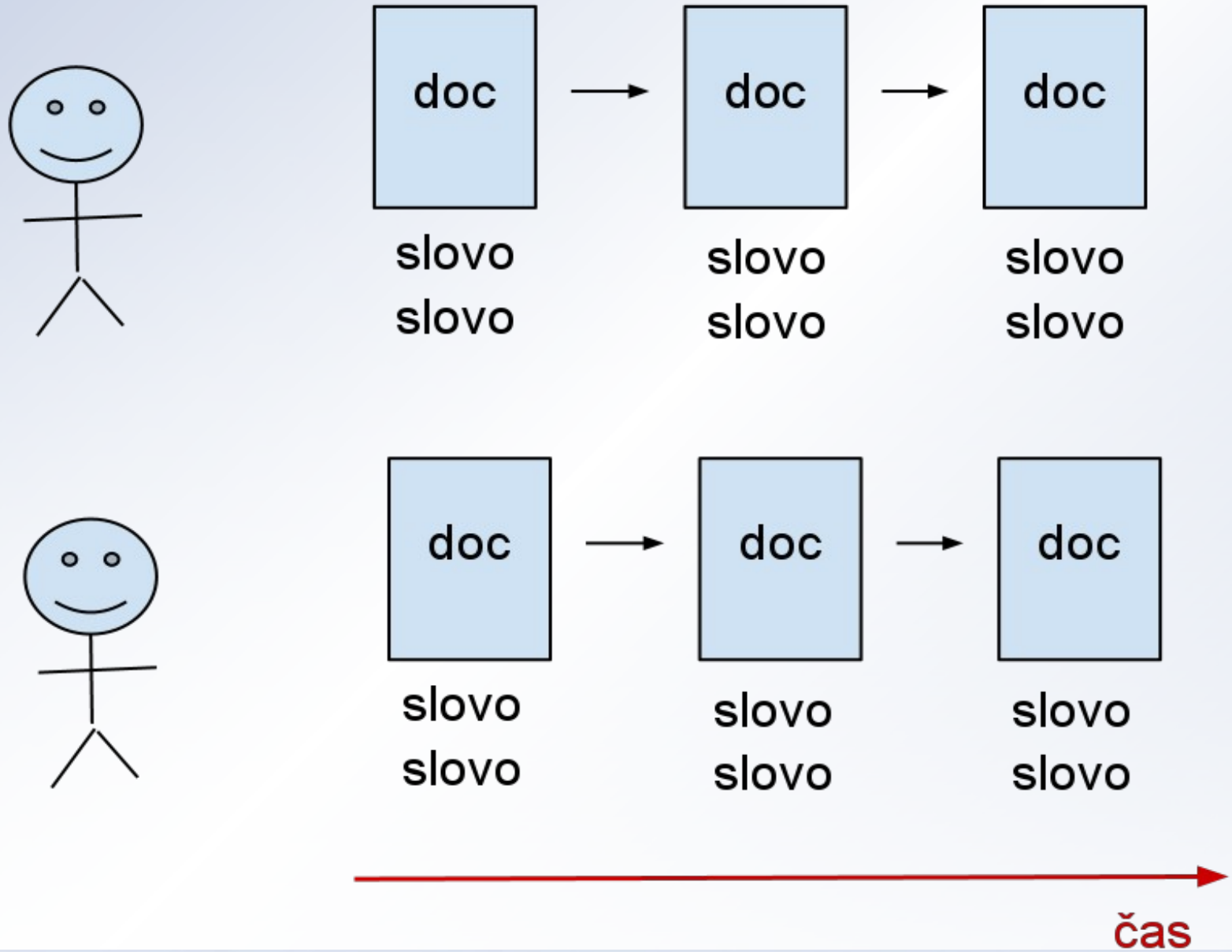
Predošlé prístupy



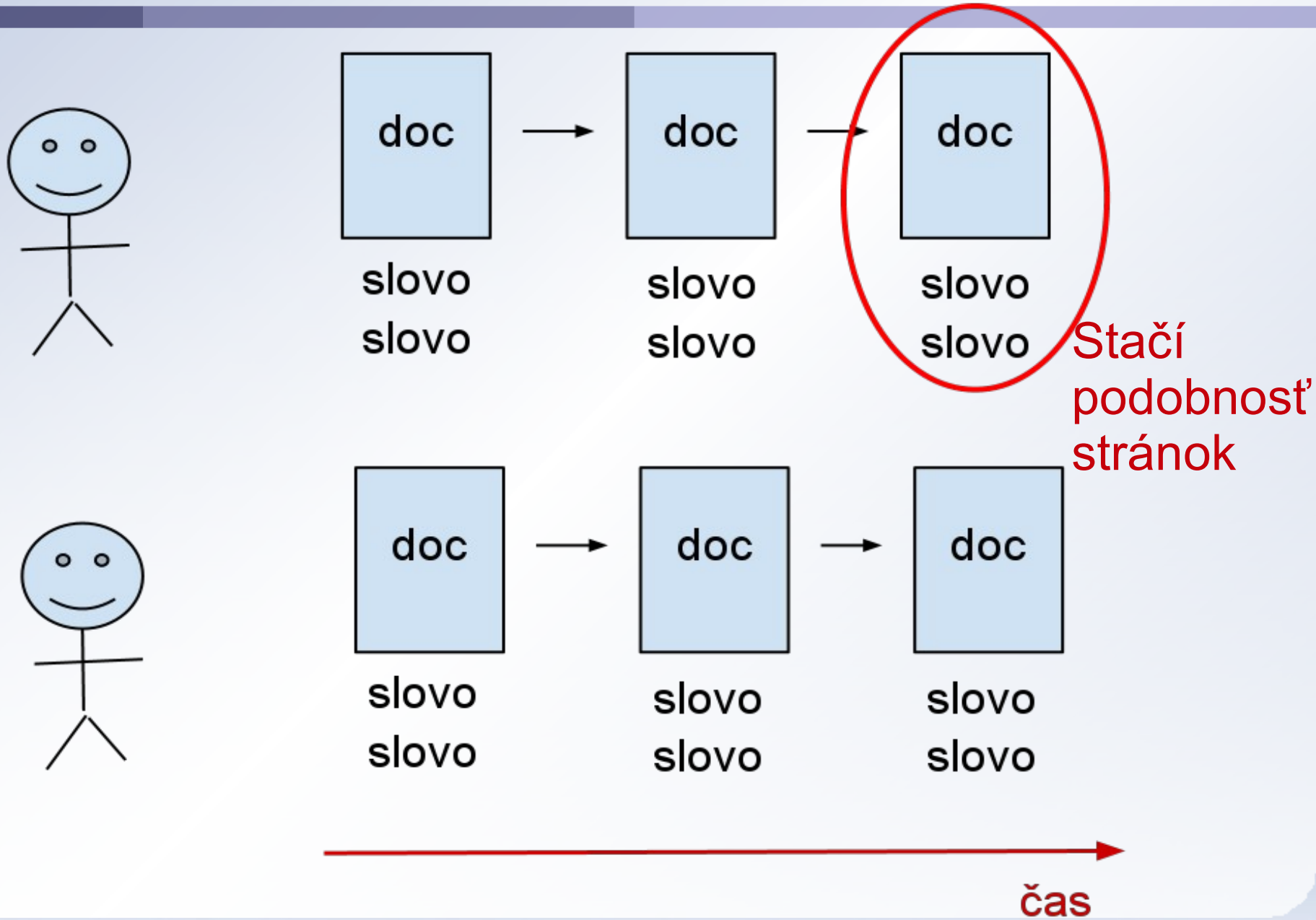
Predošlé prístupy



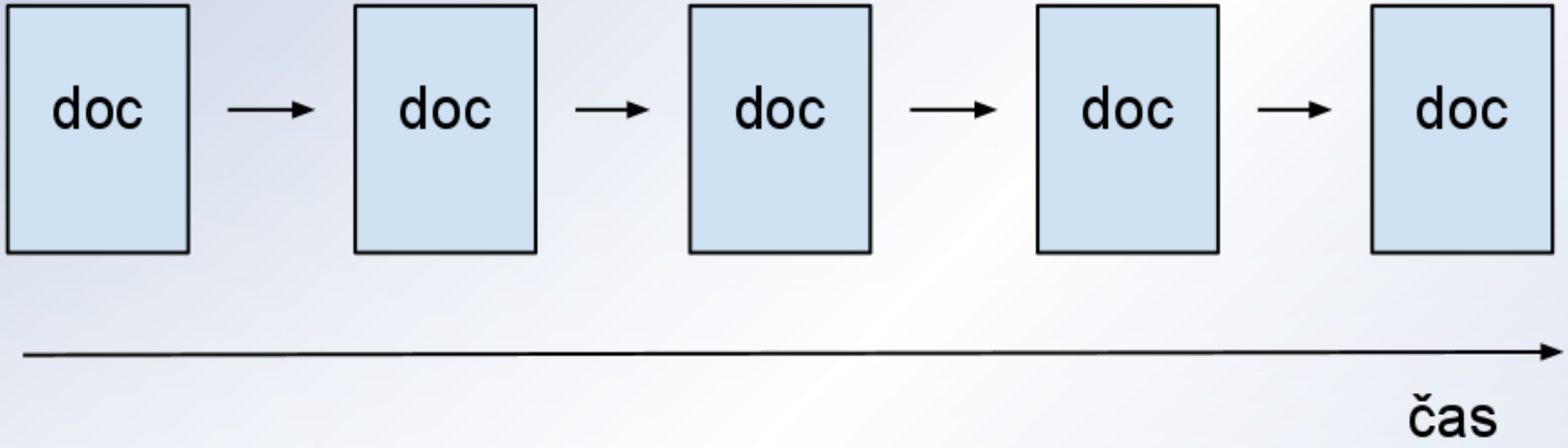
Náš prístup


doc
slovo
slovo
doc
slovo
slovo
doc
slovo
slovo
doc
slovo
slovo
doc
slovo
slovo
doc
slovo
slovo
čas

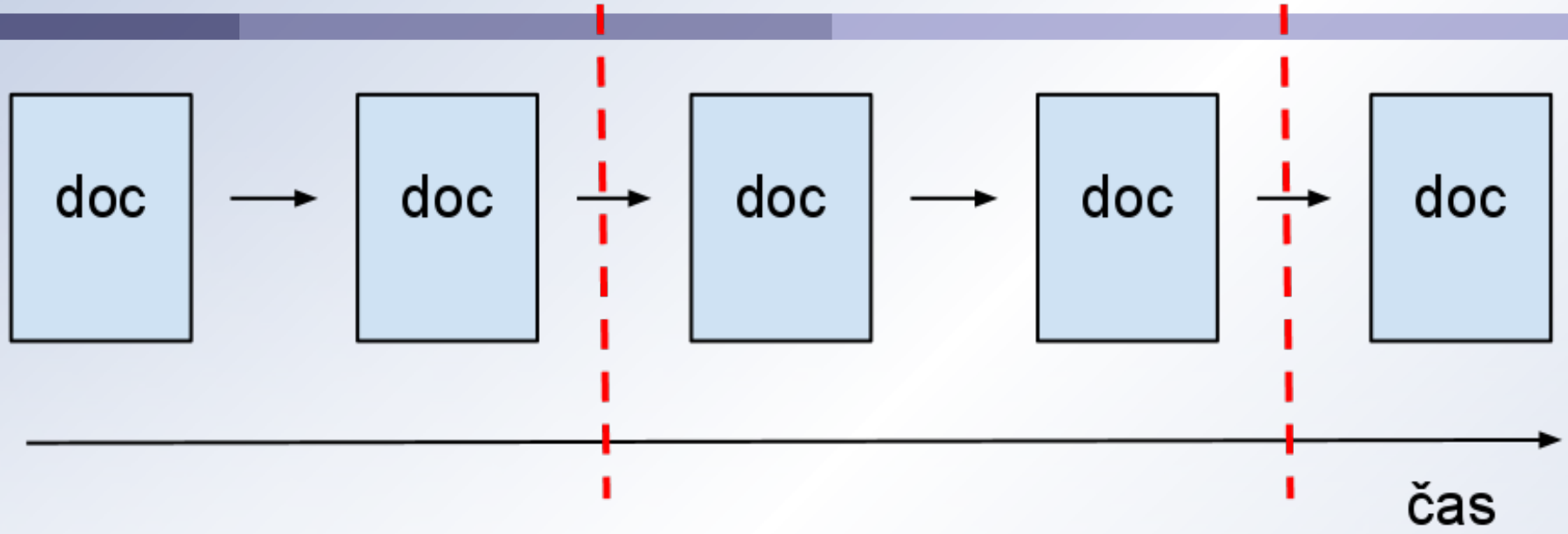
Náš prístup


doc
slovo
slovo
doc
slovo
slovo
doc
slovo
slovo
Stačí podobnosť stránok
doc
slovo
slovo
doc
slovo
slovo
doc
slovo
slovo
čas

Sessions

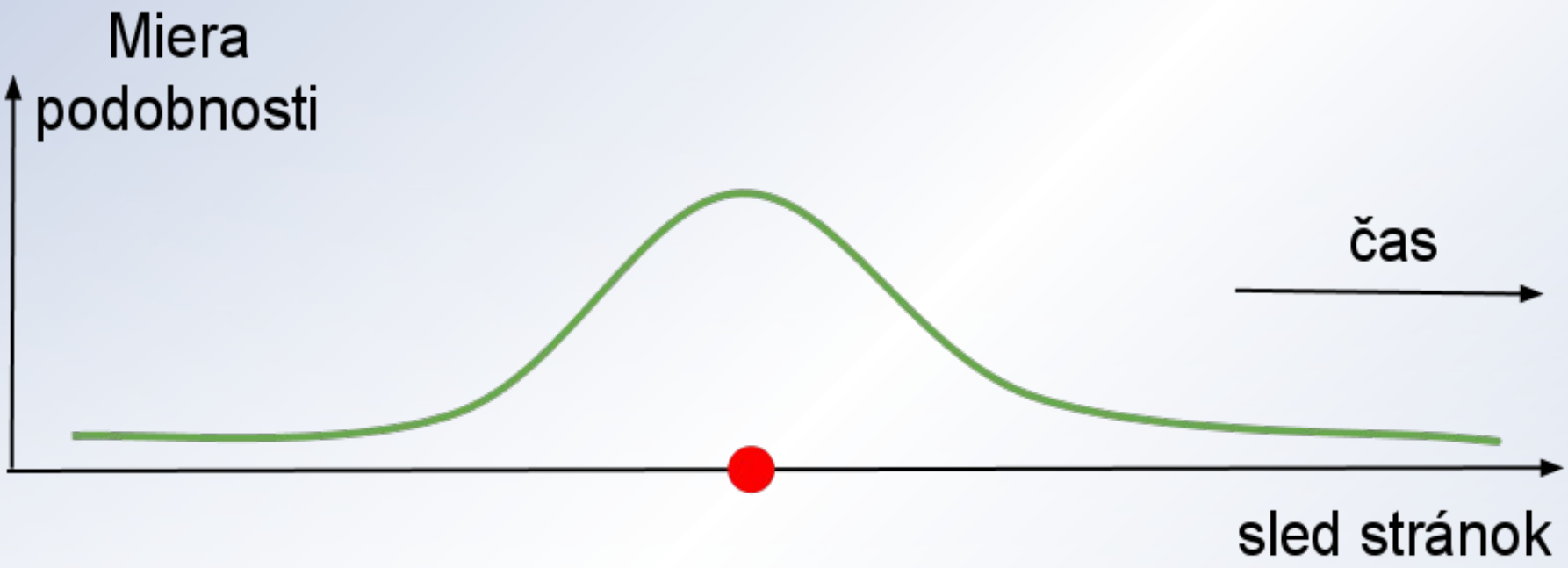


Sessions



- Časové okno
- Levensteinova vzdialenosť kľúč. slov
- Taxonómia (existujúca, Linked data)
- Naša taxonómia

Podobnosť stránok vrámci session



- Zohľadniť čas strávený na stránke
- Penalizovať časté stránky (TF-IDF)

Podobnosť kľúčových slov

- Podobnosť

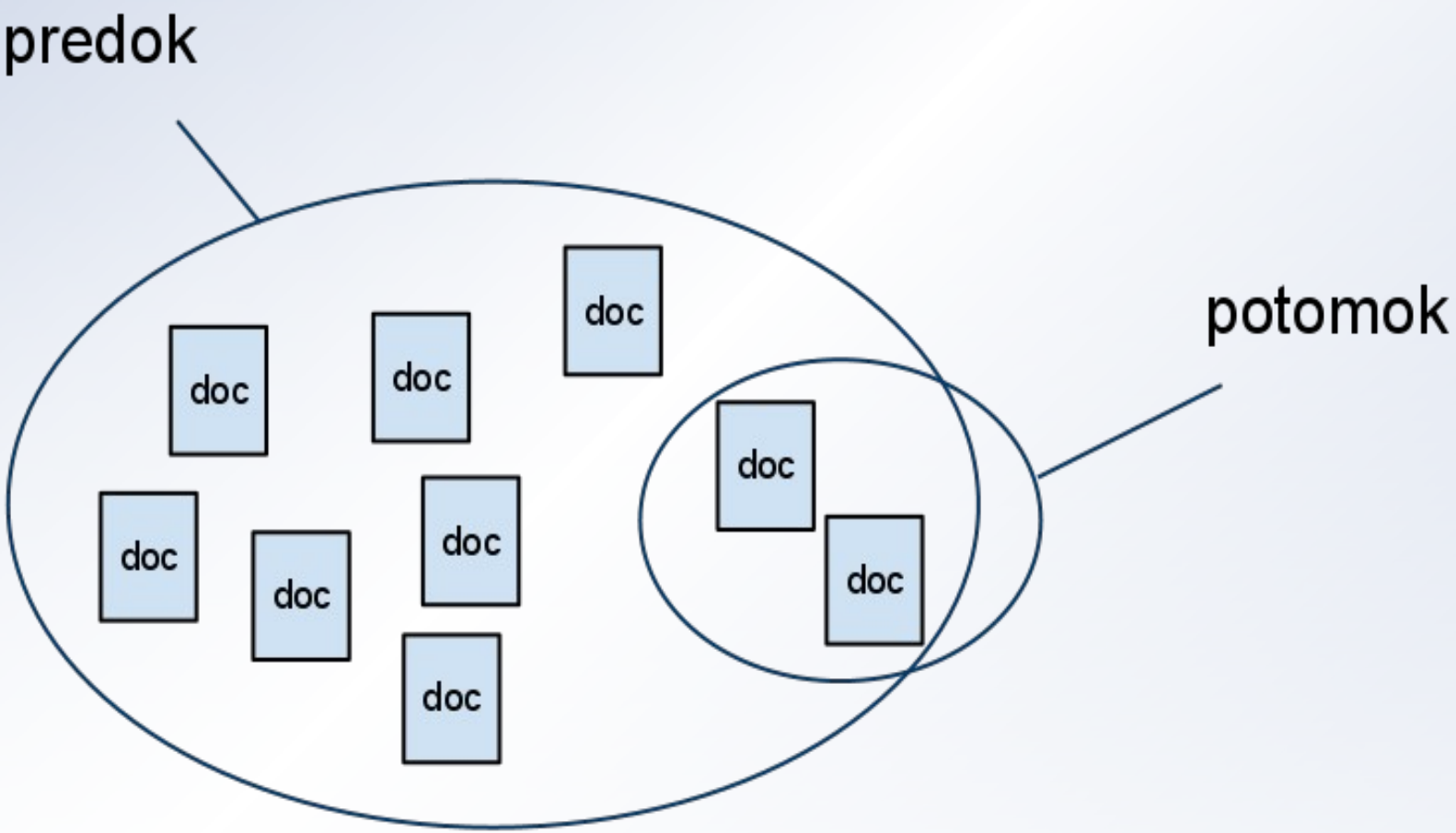
$$termSim(t_1, t_2) += pageSim(p_1, p_2) \quad \forall\, t_1 \in p_1, t_2 \in p_2, pageSim(p_1, p_2)$$

- Príslušnosť

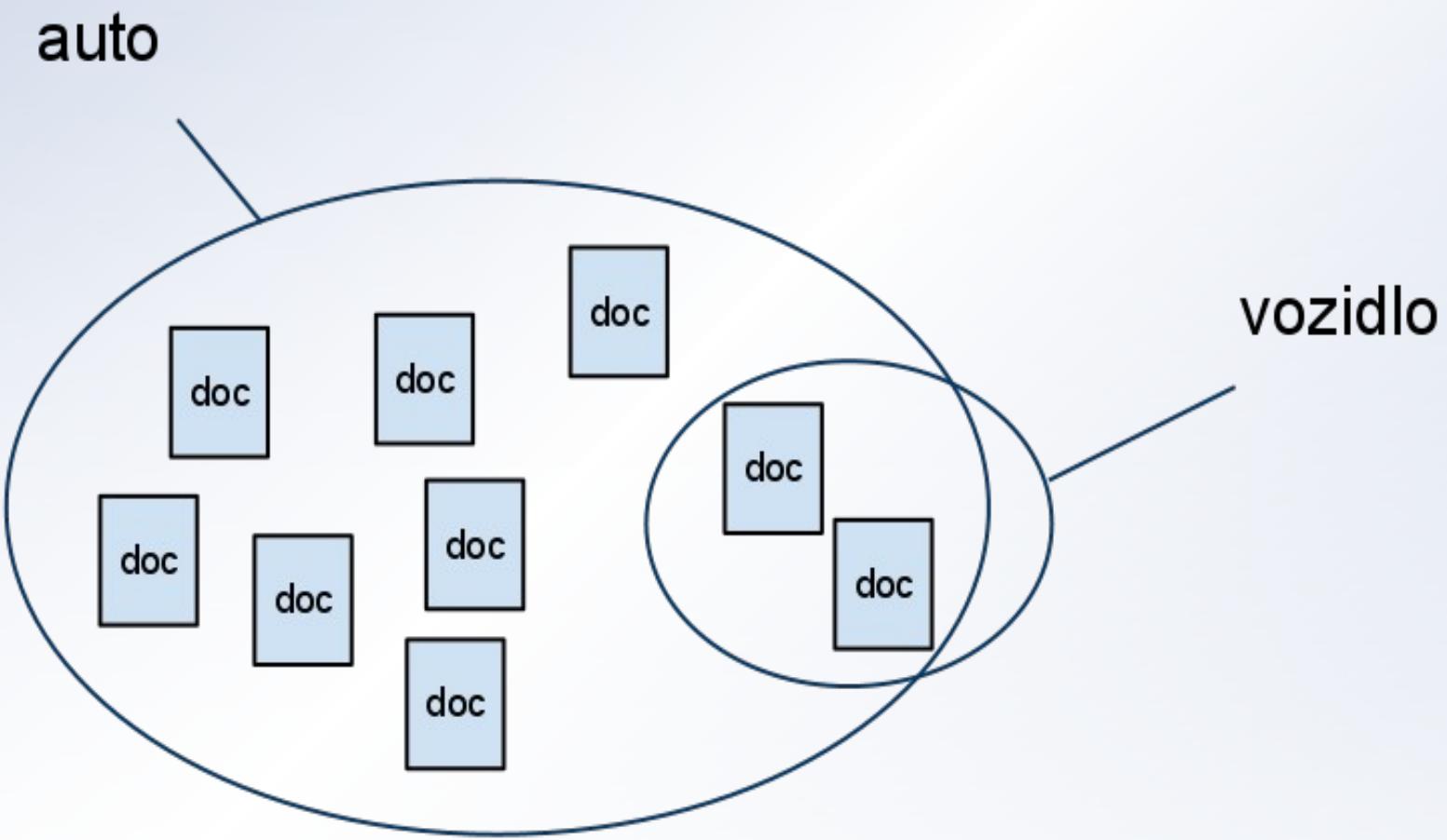
$$termSim(t_1, t_2) += tfIdf(t_1) * pageSim(p_1, p_2)$$

- Podobnosť z príslušností
- Rovnako pre stránky

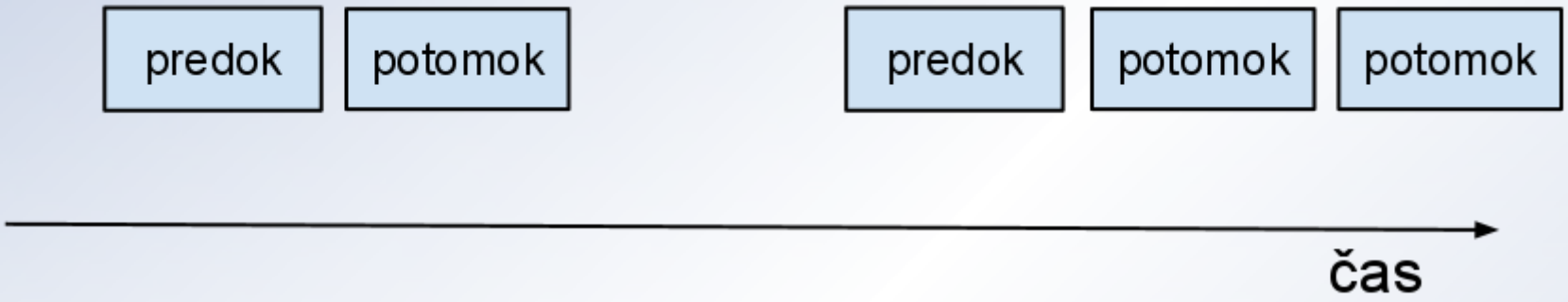
Hierarchia - predošlé prístupy



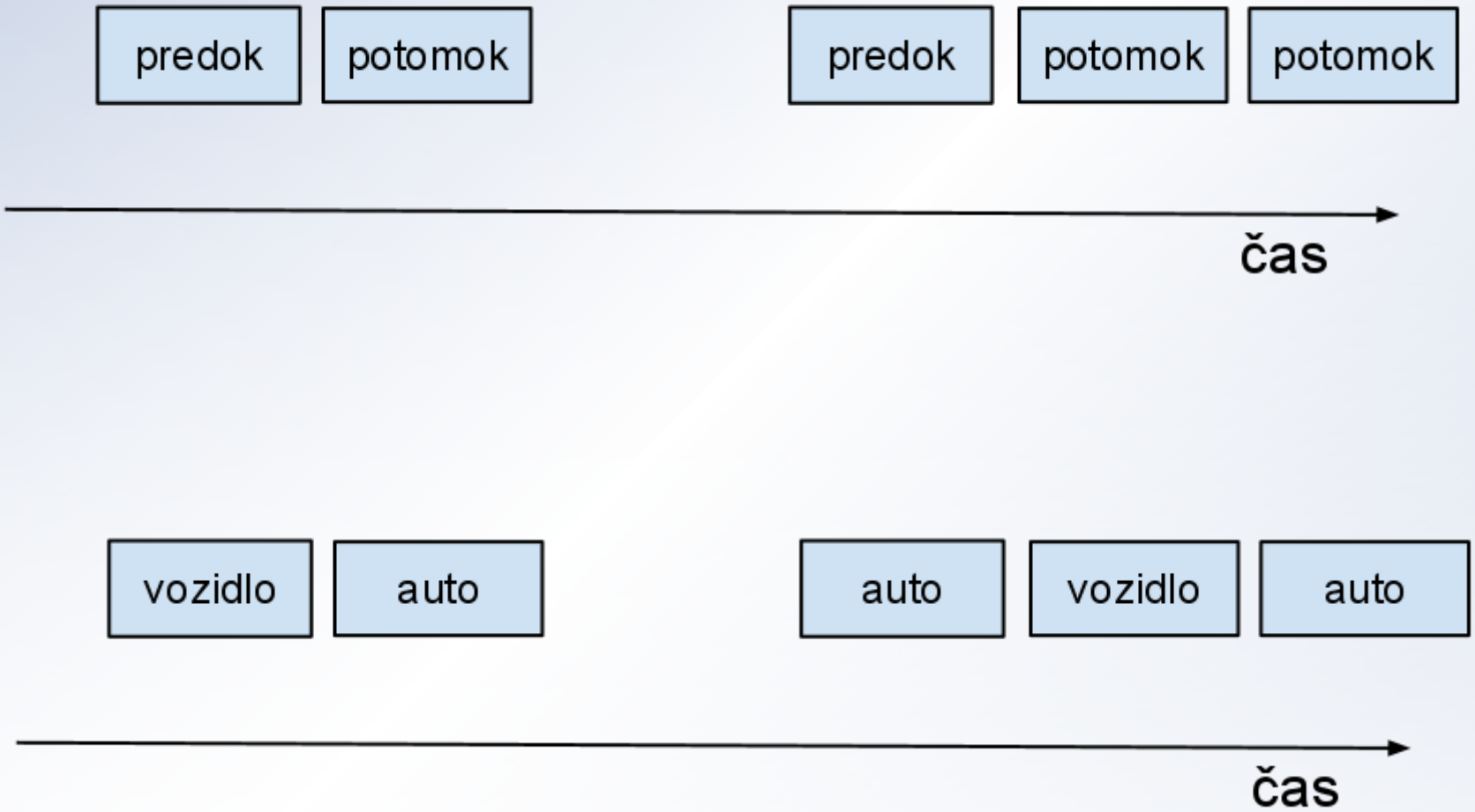
Hierarchia - predošlé prístupy



Hierarchia – náš nápad



Hierarchia – náš nápad



Výstup - klastre

session, google, field, app, name, type, Fabulous, Browse, tab session, firefox 5.0 ca, Fabulous Tabs, awesome tab, Awesome Bar, open tabs, panorama, Awesome Bar, tag

Javascript, Firefox, input, menu bar, web development platform, Vim text editor, Firefox extension, default configuration, error messages, syntax error, XPI

proxy servers, apache server

table, script, hash, cells

Repository, repositories, code, flash player, GitHub, Metal history, interactive overview, genres, Synchronize, sync button

Overenie

- Príbuznosť dokumentov
- Rozširovanie dopytov
- Odporúčanie tagov

Zhrnutie

- Vzťahy kľúčových slov z postupností
  - Viac dát, netreba tagovať
  - Časová informácia
- Detekcia sessions z taxonómie (iteratívne vylepšovanie)
- Využitie TF-IDF a asymetrickej podobnosti slov, stránok
- Identifikácia hierarchie z postupností
- Overenie riešenia