SEZNAM.CZ

Roman Weisser

MetaSearch (MS)

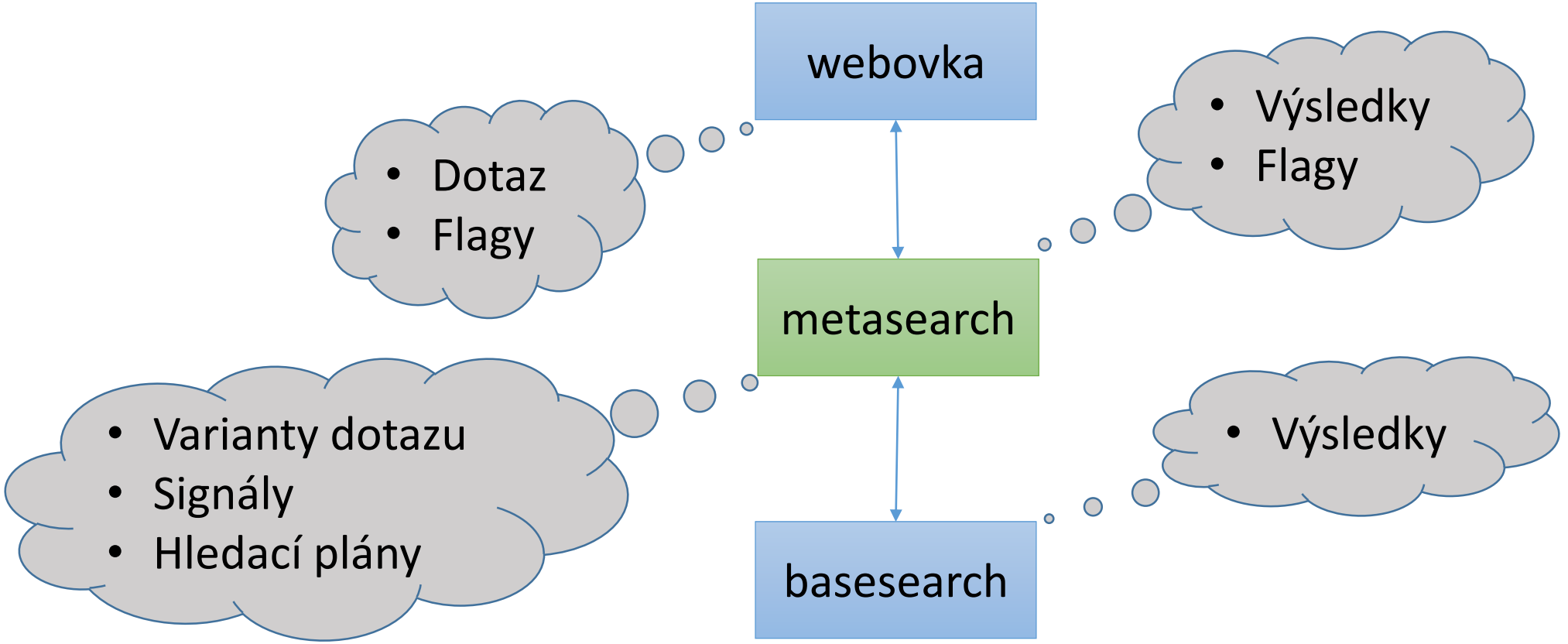
- Získává informace
- Expanduje
- Inicializuje hledání
- Předává výsledky webovce

10ms

200ms

500 dotazů/s (80 dotazů/s na MS)
15mil dotazů/den (3.5mil unikátních dotazů/den)

Komunikace a Data



Rozvoj dotazu (1/2)

Typ hran:

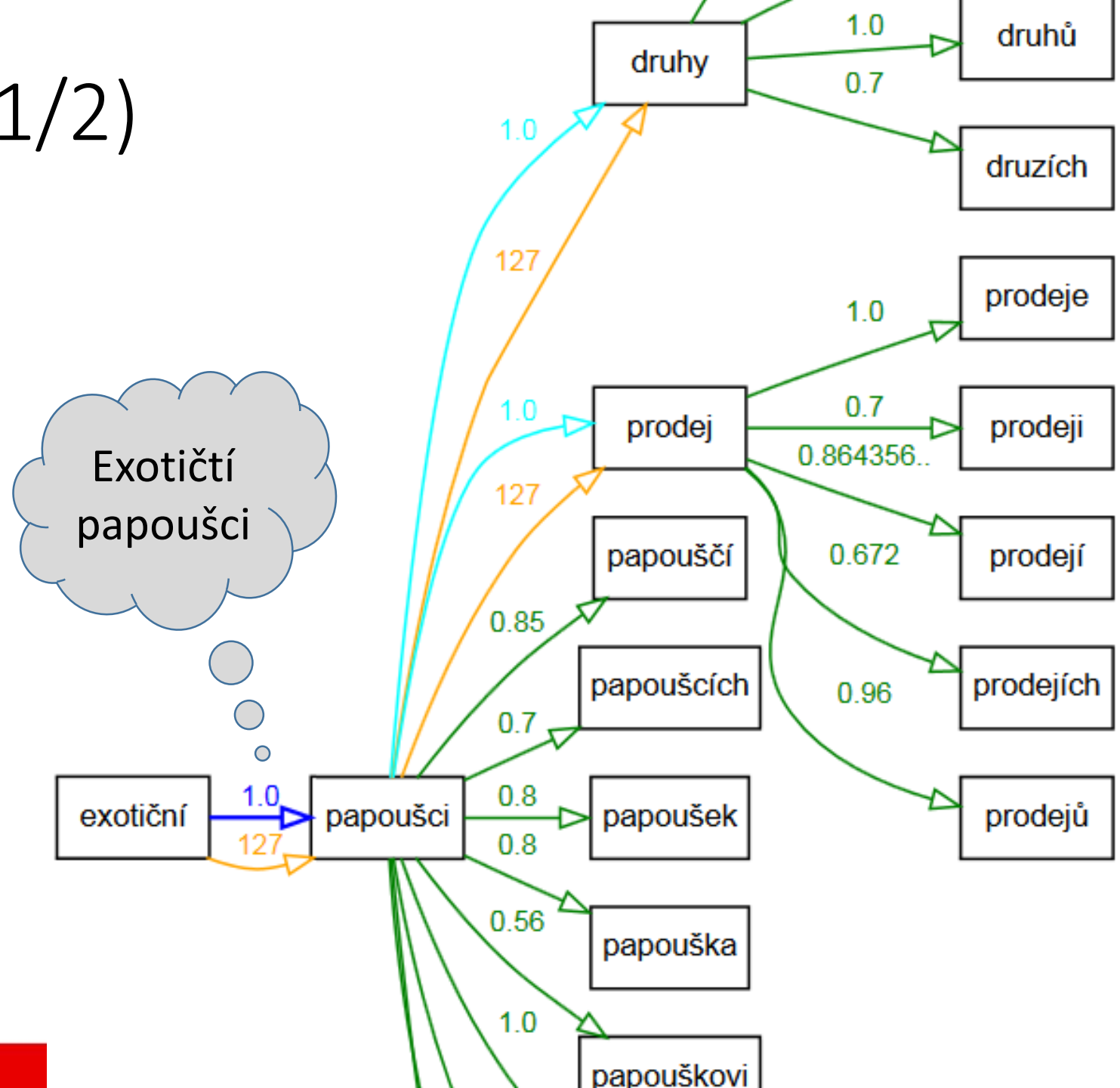
OR: zelená

AND: modrá

RELACE: žlutá

VARIANTA: červená

NEPOVINNÁ: světle modrá



Rozvoj dotazu (2/2)

Typ hran:

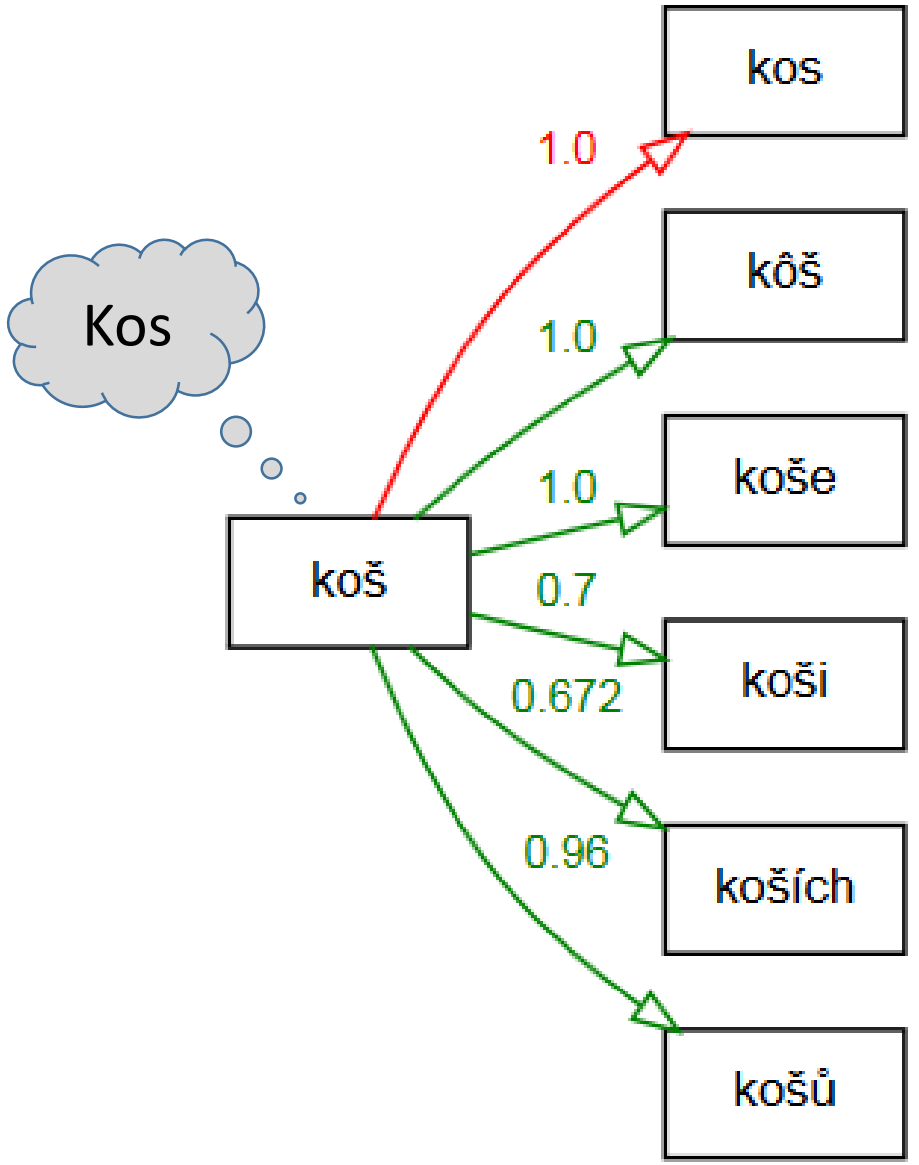
OR: zelená

AND: modrá

RELACE: žlutá

VARIANTA: červená

NEPOVINNÁ: světle modrá



Informace o dotazu

- počet slov / znaků
- četnosti slov v (korpusu, dokumentech, odkazech, dotazech)
- typy entit (jména, města, filmy, celebrity, …., kolokabilita)….

- původ (našeptávač, oprava překlepů)
- skokanovitost
- Hledanost
- …

Řešené problémy (1/2) – **Pochopení dotazu**

- Oháčkování
  - líska / liška
- Detekce jazyka
  - cap (en) / čáp (cs)
- Desambiguace lemmat
  - liska: strom / zvíře / houba / herec / čmss
- Čísla a sekvence znaků
  - 123.456.789: telefonní číslo
  - 6.5 16 5 112 57 ET43
- Odvozená slova
  - Motol / Motolská
- Alternativy
  - univerzita / universita
  - babybox / baby box
- Synonyma
  - zubař / stomatolog

Řešené problémy (2/2) – **Pochopení dotazu**

- Zkratky
  - PPC: PayPerClick / PowerPC / Pocket PC
  - Arithmetic and logic unit: ALU
- Skloňování
- Reformulace dotazu
  - Restaurace v …
  - Jak se …
- Vážení a Ignorování slov
  - dlouhé / drahé / neexistující
- Vztahy mezi slovy
  - Kolokace / pravidla
- Překlepy
  - facebook.cz / facebook.com
- Záměr (navigační, komerční, …)
- Detekce operátorů
- Detekce porno / neporno dotazů
  - velcí ptáci

Děkuji za pozornost

Otázky?

Roman Weisser
Product manager
roman.weisser@firma.seznam.cz