RSS čtečka s inteligentní filtrací

Projekt z umělé inteligence

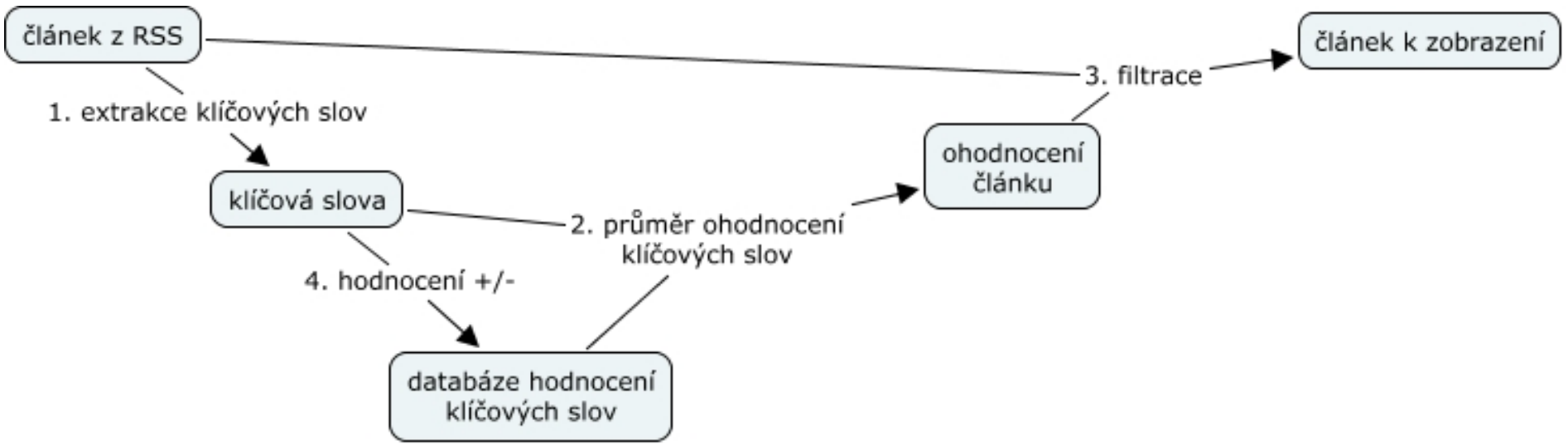
Jiří Procházka

FI MUNI

8. března 2013

Koncept

- Problém moderní informační společnosti je přehlcenost informacemi
- Mnoho novinových článků je pro uživatele nezajímavých, proto by bylo vhodné aby RSS čtečka na základě jeho preferencí články filtrovala



Stávající řešení

$\varnothing$?

Většina podobných projektů buď filtruje:

1. Pomocí ručně zadaných filtrů (např. regulární výrazy)
2. Ručním zařazením do taxonomie administrátory RSS agregátorů
3. Neznámým proprietárním způsobem

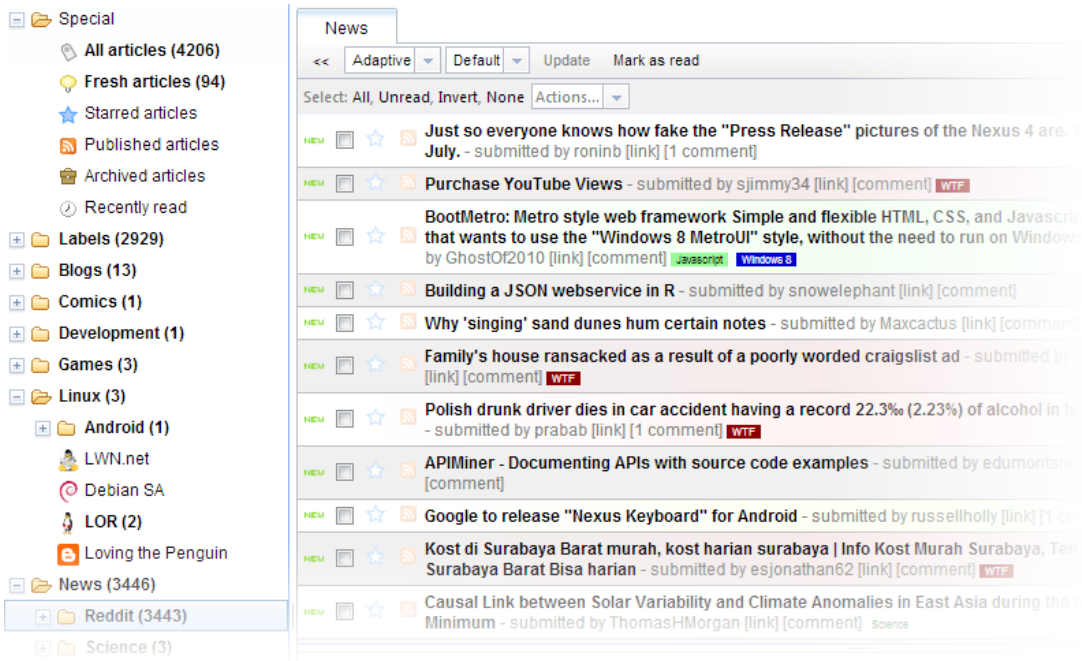
Implementace

Python

- flexibilní mnohostranný jazyk
- nltk - Natural Language Toolkit

Tiny Tiny RSS

- vyspělá RSS čtečka
- projekt koncipovat jako propojený s ní (REST interface)



Extrakce klíčových slov

- Je třeba získat strukturovanou sumarizaci článku pro ohodnocení, což budou klíčová slova.
- Protože projekt by měl fungovat na generickém obsahu, nelze předem vydedukovat cílové charakteristiky a podle toho navrhnout specifičtější strukturu a způsob extrakce informací.
- S využitím knihovny nltk lze jednoduše najít vhodné fráze jako kolokace (slova která se často vedle sebe vyskytují) na základě jejich distribuce pravděpodobnosti v obecném korpusu (což je postup pro získávání klíčových frází), nebo najít slova z článku jejichž rozdíl frekvence výskytu v článku a korpusu je velký.
  Také lze oba postupy kombinovat, a kombinovat extrakci z článku a nadpisu.

Průměr ohodnocení klíčových slov

V databázi se vybere (pokud existuje) ohodnocení extrahovaných klíčových z minulých hodnocení uživatelem a jejich průměr se použije k filtraci.

- Je také možné využít váženého průměru, na základě hodnot z extrakce klíčových slov.
- Využitím ontologie Lexvo, která vychází mimojiné z Wordnetu, lze využít informací o slovech, typu synonyma, hyponym a hypernym ("rudá" hypernymum: "červená"), a tím rozšířit množinu extrahovaných klíčových slov a zvětšit tak šanci, že budou nalezeny klíčová slova dříve hodnocená uživatelem, což urychlí "učení" programu preferencí uživatele.

Filtrace a hodnocení

Filtrace probíhá porovnáním průměru ohodnocení s uživatelem nastavenou prahovou hodnotou, po kterém čtečka článek buď zahodí, přesune do jiné složky, nebo zvýrazní jako doporučný.

Hodnocení by mělo být tlačítky + a - a možné jak u klíčových slov jednotlivě, nebo celého článku, rozdělením hodnocení klíčovým slovům.

Konec prezentace

Děkuji za pozornost