Г.С. Краліна

*Національний авіаційний університет*

МЕТОДИ ПОРІВНЯЛЬНОГО АНАЛІЗУ СУЧАСНИХ
ПОШУКОВИХ СИСТЕМ

Розглянуто автоматичні методи порівняння кількісних і якісних характеристик російськомовної частини індексів сучасних пошукових систем.

Досить актуальними представляються статистичні дослідження, присвячені розробці методів найкращого семплювання веба і наближеній оцінці відносних розмірів баз даних сучасних пошукових систем. А також визначення обсягу статичного загальнодоступного (не схованого за пошуковими формами запитів до онлайновым баз даних і не потребуючого привілейованого доступу) Інтернету. Одні з перших найбільш значимих результатів у цій області були отримані в роботах К. Бхарата та А. Бродера з Центра системних досліджень. Основна ідея їхнього підходу полягала в тому, щоб узявши за основу велику вибірку (порядку 300 тис.) документів з каталогу Yahoo!, скласти по ній представницький частотний словник (більш 400 тис. слів) максимально що близько відбиває як лексичний склад, так і реальні кількісні пропорції між словами в Інтернеті. Автори виходили з припущення, що всі пошукові системи повинні використовувати Yahoo! як один з основних інформаційних ресурсів при створенні алгоритмів обходу веба. На жаль, правдоподібність такого припущення не перевірялася.

Новизна даної роботи полягає в тому, що в ній систематично розглянуті не тільки кількісні, але і якісні параметри архівів найбільш великих пошукових систем, таких як Яндекс, Рамблер, Google і Yahoo! Крім того, розроблені методи, на відміну від закордонних аналогів, дозволяють робити вимір вебу на регулярній основі, що особливо важливо, з огляду на високу динаміку зміни стану сучасного Інтернету.