

SEO-АНАЛІЗ ТЕКСТУ

За обсягами даних World Wide Web і безліч інших систем передачі даних в мережі Інтернет є самим великим з коли-небудь існували в історії людства сховищ даних. Головне завдання пошукових систем - надавати користувачам сторінки, які максимально відповідають їхнім запитам. Головна проблема власників сайтів, які хочуть пробитися в топ видачі - створювати контент, який точно "догодить" пошуковцям.

Багато компаній все більше і більше цікавляться методами аналізу текстів. Це можуть бути запити клієнтів, внутрішнє листування, чати з клієнтами, внутрішня база знань, огляд преси або дані парсинга соцмереж. Все це ставить на перше місце питання, яким чином можна аналізувати в автоматичному режимі великі обсяги даних і надалі використовувати отримані дані.

Пошукова оптимізація (англ. Search engine optimization, SEO) - комплекс заходів щодо внутрішньої і зовнішньої оптимізації, для підняття позицій сайту в результатах видачі пошукових систем по певних запитах користувачів, з метою збільшення трафіку і потенційних клієнтів (для комерційних ресурсів) і подальшої монетизації цього трафіку.

Переваги SEO:

- якісний і актуальний контент забезпечить сайту високе місце в пошуковій видачі і стабільний приплив трафіку на сайт;
- створення якісного контенту забезпечує не тільки SEO-оптимізацію сайту, але і розширення присутності в соціальних мережах і social media;
- просування сайту по низькочастотних запитах призводить на сайт цільової трафік, який можна конвертувати, пропонуючи пропозиції;
- SEO-оптимізація спрямована на підвищення позицій сайту в пошукових системах і репутації сайту за рахунок посилань на сайт на зовнішніх джерелах;
- SEO-технології забезпечують корисність сайту для відвідувачів.

Недоліки SEO:

- створення та адміністрування якісного контенту вимагає багато часу і ресурсів (фінансових, трудових та ін.);
- для ефективної SEO-оптимізації сайту крім створення якісного контенту необхідні навички роботи з технічними аспектами оптимізації (редирект, робота з Хостін-гами і ін.), що вимагає залучення веб-розробників;

- у різних пошукових системах переважають різних чинники ранжирування, тому необхідно розробляти кілька SEO-стратегій;
- необхідно постійно проводити моніторинг сайту і використовувати веб-аналітику.

На жаль, більшість SEO-оптимізаторів не використовують статистичну близькість між своєю оптимізацією (входами) та відповідними їй пошуковими результатами (виходами). Більш того, їх входи або оптимізації знаходяться перед компонентами обходу в пошуковій системі. Фактично типова оптимізація повинна пройти через 4 рівня: обхід вмісту, індексування, скорингу та, в кінцевому підсумку, рівень запиту в реальному часі.

Для вирішення оптимізаційних задач використовується поєднання семантичних і математичних моделей. Семантична модель використовується для того, щоб відобразити все багатство атрибутів і зв'язків елементів системи, що моделюється. На моделі проводяться логічні і математичні обчислення, які дозволяють знайти оптимальне рішення. Як правило, таке моделювання є козним: модель постійно оновлюється за рахунок надходження нових відомостей про стан системи, що моделюється, і, відповідно, автоматично перераховується оптимальне рішення, що призводить до корекції програми дій.

Серед величезного числа алгоритмів, які використовуються для пошуку і аналізу інформації, особливе місце займають ті, чия мета - виявлення прихованих закономірностей або неочевидних залежностей.

Про один з методів, який застосовується для рекомендаційних систем (коллаборативна фільтрація), інформаційного семантичного пошуку, поділу текстів по тематиках без навчання та багатьох інших і піде мова далі. Метод цей називається латентно-семантичним аналізом (LSA - Latent semantic analysis). Можна сказати, що це просунутий SEO аналіз тексту.

Його суть полягає в тому, що для нас не важлива порядок слів у документі, в яких морфологічних формах вони представлені, а важливо тільки кількість входжень конкретних слів.

На обґрунтуванні тільки цієї інформації, будується таблиця «слово-документ». Де рядки відповідають словам (а точніше, їх леми), а стовпці - документам. У кожному осередку зберігається 1, якщо слово є в документі, і 0 - якщо ні. Хоча такий варіант і найпростіший, але не найкращий. Замість 0 і 1 можна використовувати, наприклад, частоту слова в документі або tf-idf слова. Такий спосіб представлення текстів у вигляді таблиці (або матриці) називається векторною моделлю тексту. Тепер, для того щоб порівняти два документи, потрібно визначити міру схожості двох стовпців таблиці.

Висновки. LSA дозволяє підтримувати SEO-оптимізацію контенту. Після використання LSA ми отримуємо семантичне уявлення слів і документів - це дозволяє знаходити неочевидні зв'язки між словами і

документами. До недоліків алгоритму можна віднести дуже велику обчислювальну складність методу.