УДК 004.89

Дудар В. – ст. гр. СНм – 51
Тернопільський національний технічний університет імені Івана Пулюя

МЕХАНІЗМИ ПОШУКОВИХ СИСТЕМ

Науковий керівник: асистент Маєвський О. В.

Пошукові системи зазвичай складаються з трьох компонентів: агент (павук або кроулер), який переміщається по мережі і збирає інформацію; база даних, яка містить усю інформацію, що збирається павуками; пошуковий механізм, який люди використовують як інтерфейс для взаємодії з базою даних.

Засоби пошуку типу агентів, павуків, кроулерів і роботів використовуються для збору інформації про документи, що знаходяться в мережі Інтернет.

Кожен пошуковий механізм має власний набір правил, які визначають, як збирати документи. Деякі йдуть за кожним посиланням на кожній знайденій сторінці і потім, у свою чергу, досліджують кожне посилання на кожній з нових сторінок, і так далі. Деякі ігнорують посилання, які ведуть до графічних і мультимедійних файлів; інші ігнорують посилання до ресурсів типу баз даних WAIS; інші проінструктовані, що треба переглядати передусім найбільш популярні сторінки.

Самими «інтелектуальними» з пошукових систем є агенти. Вони можуть робити більше, ніж просто шукати: вони можуть виконувати навіть транзакції від Вашого імені. Вже зараз вони можуть шукати сайти специфічної тематики і повертати списки сайтів, відсортованих по їх відвідуваності. Агенти можуть обробляти зміст документів, знаходити і індексувати інші види ресурсів, не лише сторінки. Вони можуть також бути запрограмовані для видобування інформації із вже існуючих баз даних. Незалежно від інформації, яку агенти індексують, вони передають її назад базі даних пошукового механізму.

Загальний пошук інформації в мережі здійснюють програми, відомі як павуки. Павуки повідомляють про зміст знайденого документу, індексують його і видобувають підсумкову інформацію. Також вони переглядають заголовки, деякі посилання і посилають проіндексовану інформацію базі даних пошукового механізму.

Кроулери переглядають заголовки і повертають тільки перше посилання.

Роботи можуть бути запрограмовані так, щоб переходити по різних посиланнях різної глибини вкладеності, виконувати індексацію і навіть перевіряти посилання в документі. Через їх природу вони можуть зациклюватись, тому, проходячи по посиланнях, їм потрібні значні ресурси мережі. Проте, є методи, призначені для того, щоб заборонити роботам пошук по сайтах, власники яких не бажають, щоб вони були проіндексовані.

База даних відшуковує предмет запиту, заснований на інформації, вказаній в заповненій формі, і виводить відповідні документи, підготовлені базою даних.

Щоб визначити порядок, в якому список документів буде показаний, база даних застосовує алгоритм ранжування. В ідеальному випадку, документи, найбільш релевантні призначеному для користувача запиту будуть поміщені першими в списку. Різні пошукові системи використовують різні алгоритми ранжування, проте.

Різні пошукові механізми також вибирають різні способи показу отриманого списку – деякі показують тільки посилання; інші виводять посилання з першими декількома пропозиціями, що містяться в документі або заголовок документу разом з посиланням.