Вопрос №2.16

Архитектура поисковых систем. PageRank. Автоматическая классификация текстов. Семантические сети, Semantic web. Оценка качества результатов поиска.

1 Архитектура поисковых систем

- Поисковый робот (web crawler). Ходит по ссылкам в Интернете и индексирует найденные документы, занося в базу данных.
- База данных. Накапливает информацию о найденных страницах в формате, удобном для использования алгоритмами поиска: прямой (документ — слова) и обратный (слово — документы) индексы.
- Клиент. Обрабатывает поисковые запросы пользователей, извлекая информацию из базы данных.

Альтернативные архитектуры.

- Распределенная поисковая система: база данных распределена по ряду серверов.
- Метапоисковая система: вместо собственного поискового робота и базы данных используются запросы к другим поисковым системам.

2 PageRank

PageRank — алгоритм ранжирования результатов поиска. Ранжирование производится по рейтингу, вычисляемому через количество ссылок на данный документ и рейтинги ссылающихся документов.

PageRank использует модель случайного блуждания.

- Ориентированный граф: документы — вершины, ссылки — дуги.
- Начинаем в случайной вершине.
- С вероятностью ε переходим в случайную вершину.
- С вероятностью $1 - \varepsilon$ переходим по случайной исходящей дуге.

$\mathrm{PR}_k(i)$ — вероятность находиться в вершине i на k-м шаге. $\mathrm{PR}_0(i) = 1/n$, где n — количество вершин.

$\mathrm{PR}(i) = \lim_{k\to\infty} \mathrm{PR}_k(i)$. На практике ограничиваются $\mathrm{PR}_{50}(i)$.

Важные соотношения:

$$\mathrm{PR}_k(i) = \varepsilon + (1-\varepsilon) \sum_{j:\, j\to i} \frac{\mathrm{PR}_{k-1}(j)}{|\{l : j \to l\}|},$$

$$\lim_{k\to\infty}: \quad \mathrm{PR}(i) = \varepsilon + (1-\varepsilon) \sum_{j:\, j\to i} \frac{\mathrm{PR}(j)}{|\{l : j \to l\}|},$$

где $x \to y$ означает наличие дуги из x в y.

PageRank в векторной форме.

- $\overline{\mathrm{PR}}_k = (\mathrm{PR}_k(1), \ldots, \mathrm{PR}_k(n))^T$
- $L = (l_{ij})$:
  - $l_{ij} = \varepsilon + (1-\varepsilon)\frac{1}{|\{l:j\to l\}|}$, если $j \to i$,
  - $l_{ij} = \varepsilon$ иначе.
- $\overline{\mathrm{PR}}_k = L^k \overline{\mathrm{PR}}_0$
- $\overline{\mathrm{PR}} = L\overline{\mathrm{PR}}$, т.е. $\overline{\mathrm{PR}}$ — собственный вектор матрицы L.

3 Автоматическая классификация текстов.

Постановка задачи.

- Множество документов $D = \{D_1, \ldots, D_n\}$.
- Множество категорий $C = \{C_1, \ldots, C_m\}$.
- Неизвестная функция $\Phi : D \times C \to \{0, 1\}$.
- Задача: построить классификатор Φ', максимально близкий к Φ.
- Иногда достаточно построить $\Phi'' : D \times C \to [0, 1]$, который задает не точную классификацию, а ранжирование категорий для каждого документа.
- От ранжирования легко перейти к точной классификации, введя некоторый порог $0 < t < 1$: $\Phi'(D_i, C_j) = 1 \iff \Phi''(D_i, C_j) \geq t$.

Применение.

- Фильтрация документов, распознавание спама.
- Наполнение интернет-каталогов.
- Классификация новостей.
- Контекстная реклама.
- Персональные новости.

Пусть каждый документ D_i представлен в виде вектора весов термов $(t_1, \ldots, t_p)$. Рассмотрим пример линейного on-line классификатора.

- $\Phi'(D_i, C_j) = \frac{D_i \cdot C_j}{|D_i||C_j|}$, где векторы коэффициентов C_j вычисляются динамически по мере обработки обучающего множества.
- Начинаем с $C_j = (1, \ldots, 1)$.
- Для каждого учебного документа применяем текущее правило.
- При неудаче вносим поправки $+\alpha/-\beta$ в коэффициенты, соответствующие термам неправильно классифицированного документа.

4 Семантические сети, Semantic web.

Семантическая сеть (semantic network) — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги задают отношения между ними.

Семантические сети бывают однородными (только один тип отношения) и неоднородными (разные типы отношений).

Semantic Web — часть глобальной концепции развития сети Интернет, целью которой является реализация возможности машинной обработки информации за счет использования семантических сетей для описания документов и их отношений.

Преимущества Semantic Web.

- Семантический поиск.
- Вопросо-ответные системы.
- Агенты в семантическом Вебе.
- Объединение знаний (интеграция баз данных).

Языки описания для Semantic Web (основаны на XML).

- RDF (Resource Description Framework) — описывает конкретные объекты и их отношения (синтаксис).
- OWL (Ontology Web Language) — описывает типы объектов и типы отношений (семантика).

5 Оценка качества результатов поиска.

- Точность (precision) — отношение числа найденных релевантных документов к общему числу найденных документов.
- Полнота (recall) — отношение числа найденных релевантных документов к общему числу релевантных документов в базе.
- F-мера (F-measure, мера Ван Ризбергена) — среднее гармоническое точности и полноты: $F = 2PR/(P + R)$.

Основной источник информации - курс «Алгоритмы для Интернета»: `http://yury.name/internet.html`