1 Course Objectives

The objectives of the course are to present the theoretical and engineering aspects of search engine technology. The course will cover the data structures and algorithms that are in use in the major components of large scale search engines, as well as characteristics and attributes of the Web corpus and search engine users. Students of the course should exit it with an understanding of both the machinery and ecosystem of search engines.

2 Intended Audience and Prerequisites

The course's intended audience includes 4’th year BSc students and graduate students. Its prerequisites are as follows:

- Basic course in Algebra (e.g. 104167)
- Basic course in Probability (e.g. 094412)
- Basic course on Data Structures (e.g. 234218)
- Basic course on Algorithms (e.g. 234247)

While the formal prerequisites may be satisfied by 3rd year students, the course requires some mathematical and analytical maturity that 3rd year students typically lack. In particular, basic understanding of stochastic processes (e.g. as taught in 094314 or 044202) is highly recommended.

3 Grading Policy

The course consists of 13 weekly lectures. Grading will be done as follows:

- (40p) 3-4 homework assignments.
- (60p) final exam.

Wet assignments will be submitted in pairs, while dry assignments will be submitted individually.
4 Week-by-week Syllabus

<table>
<thead>
<tr>
<th>Week</th>
<th>Lecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Course outline, popular introduction to search engines, technical overview of search engine components [12, 5]; Introduction to Information Retrieval: Boolean model, vector space model, TF/IDF scoring [45, 8].</td>
</tr>
<tr>
<td>2</td>
<td>Introduction to Information Retrieval: Probabilistic models [8, 36]</td>
</tr>
<tr>
<td>3</td>
<td>Link Analysis basics: quick overview of Perron-Frobenius theory and ergodicity [22]; Google’s PageRank [12], Kleinberg’s HITS [26]</td>
</tr>
<tr>
<td>4</td>
<td>Advanced link analysis: topic-sensitive PageRank [24]; PageRank perturbations [41, 17]; SALSA and the TKC effect [31]</td>
</tr>
<tr>
<td>5</td>
<td>Web graph structure and models: power laws [39], Bow-tie structure [14], self-similarity [19]; evolutionary models of the Web graph [27, 29]</td>
</tr>
<tr>
<td>6</td>
<td>Indexing basics: what is an inverted index, how is one constructed efficiently, what operations does it support, what extra payload is usually stored in search engines, the accompanying lexicon</td>
</tr>
<tr>
<td>7</td>
<td>Query evaluation schemes: term-at-a-time vs. doc-at-a-time, result heaps, early termination/pruning, WAND [13]</td>
</tr>
<tr>
<td>8</td>
<td>Distributed index architectures: global/local schemes [5, 38, 15, 43]; the Google cluster architecture [10]</td>
</tr>
<tr>
<td>9</td>
<td>Crawlers - purpose and architecture [25, 30], optimizing crawl order [18, 46, 40], computation of importance metrics during crawl [1]</td>
</tr>
<tr>
<td>10</td>
<td>Effective caching and prefetching of query results [37, 33, 32, 7, 6, 20]</td>
</tr>
<tr>
<td>11</td>
<td>Computational advertising: online advertising ecosystem; payment models (CPM, CPC, CPA); auction mechanisms based on the pioneering course on this subject taught at Stanford University, <a href="http://www.stanford.edu/class/msande239/">http://www.stanford.edu/class/msande239/</a> and [42, 3, 21, 44]</td>
</tr>
<tr>
<td>12</td>
<td>Task Completion and Search Assistance - from spell corrections and simple shortcuts to rich media, mashups, query completions and facets [16, 11, 9, 34]</td>
</tr>
<tr>
<td>13</td>
<td>The Long Tail [4], recommender systems and collaborative filtering [2, 28, 35, 23]</td>
</tr>
</tbody>
</table>

References


