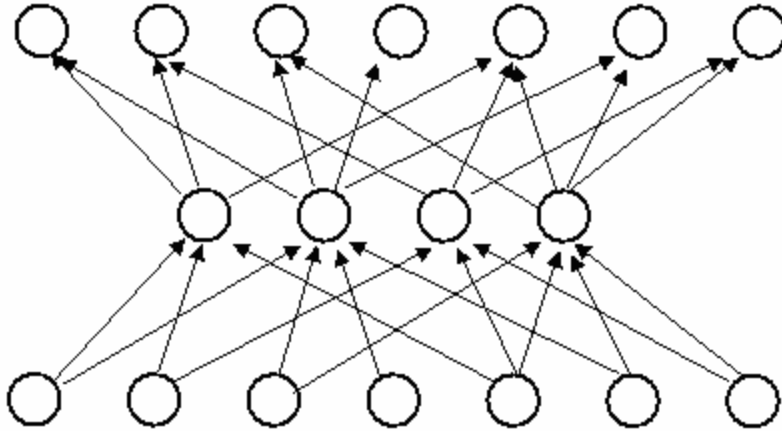


Information Retrieval Technologies In Modern Search Engines



An introduction to IR Systems in Search Engine Architecture

By Jason J. Green

Contents

- 1) Introduction
- 2) Primary Software Architecture
- 3) Indexing Process Briefly
- 4) Vector Space Model
- 5) Semantics and the Lexicon
- 6) Ranking Concepts
- 7) Link Graph Inter-Connectivity
- 8) References and Citations

Introduction

This paper is a brief exploration of various technologies and design concepts in use by modern large-scale internet search engines. No specificity is implied in regards to particular search engines, rather this document hopes to provide an overview of commonalities in modern data mining, document indexing and relevancy weighting methodologies currently employed by many commercial search engines.

Primary Software Architecture

The following sections describe software and database components common to modern search indexes and Information Retrieval systems.

Indexing Robots -

The indexer is usually a team of distributed web robots (spiders) that collect web documents. Spiders also collect information about document inter-relatedness. Web traversal is accomplished by following hyper-links.

Document Repository -

The document Repository stores indexed web pages in a compressed format. Depending on the engine itself this is either a complete document index (Such as Google, which downloads and stores complete web pages.), or a partial document index which contains condensed representations of web pages composed of those elements which the search engine uses to ascertain topical information about a web page.

Document Indexer -

The document Indexer is the sorting and organization hub of the search index. It stores information about the web pages in the document Repository which are usually sorted according to an Identification scheme specific to each search engine. The document indexer can double as a table of contents for the document repository, storing information such as URL, unique ID, document size, etc.

Inverted Index-

The Inverted Index is a specialized document Index which stores and organizes web pages by word composition information. During indexing, a web page is parsed into word counts. For example the following statement:

“See Spot run. Spot likes to run with Jack.”

Would be represented in the Inverted Index as:

See=1, Spot=2, Run=2, Likes=1, To=1, With=1, Jack=1

The inverted index can also contain information about word emphasis, capitalization, placement, etc.

Lexicon -

The Lexicon is the very backbone of the document analysis process and is basically an extremely comprehensive dictionary / thesaurus database against which documents are evaluated for word references, word usage grammar and syntax. The lexicon can be an extremely complex language model which is used to calculate document “positions” in a virtual 3D space construct (see Vector Space Model).

Indexing Process Briefly

A URL is sent to the spidering queue by the URL server and a web robot downloads the associated web page (or specific information from the web page) and sends that web page to the document repository.

The Document Repository compresses the web page for processing and storage and a unique Identification is assigned to the web page which is then sent to the document indexer.

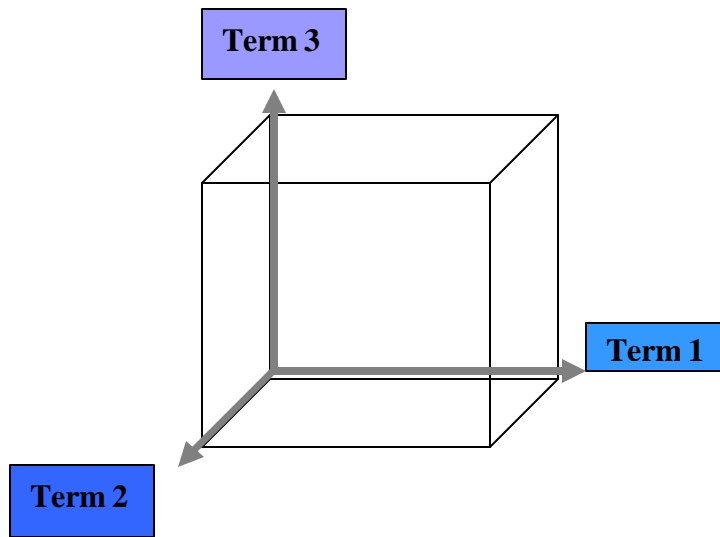
The Document Indexer sorts the unique identifier of the web page amongst the rest of the documents list and references location information so that the document can be accessed from the repository. All link information including link text is parsed from the web page and is usually stored in a separate index so that a citation ranking can be assigned to the web page.

The document index then parses the web page into word occurrences and packages this information for storage in the inverted index. Generally the document index will also store information about the web page such as size, term positioning, emphasis, etc. The web page is assigned scores or weights based upon all of the above factors according to the methodology of the particular search engine. Using the parsed and ranked term information, the document is organized into word identifications which are used to create the inverted index.

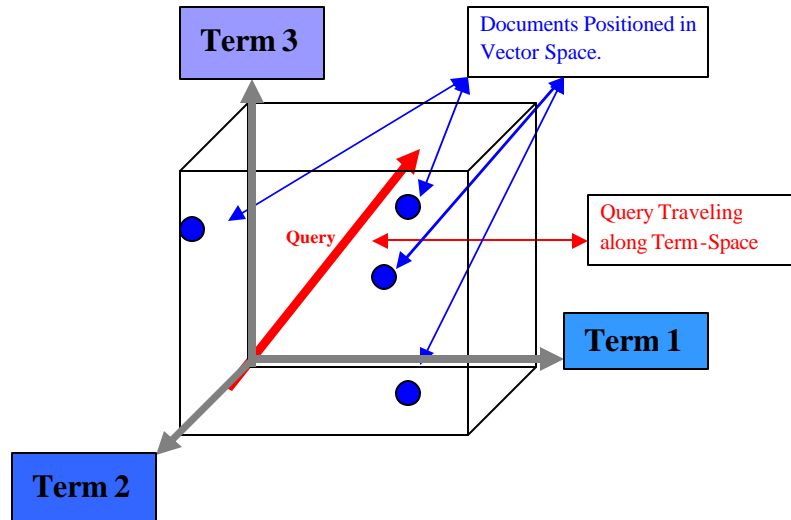
The Vector Space Model

In this section we will briefly touch upon some basic concepts of the Vector Space Model; which forms the foundation for modern IR theory and consequently search engine software architecture.

The Vector Space Model (Salton's Vector Space Model or Term Vector Model) is a multidimensional model of terms arranged according to a semantic-proximity scheme that situates lexicon terms in space according to their conceptual ontology.



Search queries are projected as vectors (implying magnitude and direction) while documents are mapped as vector points in term space according to their conceptual relationship to appropriate terms.



Queries travel along a vector which represents relation to terms while Query and Document weights are based on the length and direction of their vector.

Term Weights in Vector Space

A Term Weight is an evaluation of the relative importance of any given term in relation to other terms within a vector model. Variations of this method are commonly employed by modern search engines. A Vector Space IR system is concerned with three primary evaluations. They are:

1. The global weight of every indexed term in relation to the entire document set. (How important is this term?)
2. The document weight of every term within a given document. (How important is this term to the Document?)
3. Query term vectors are evaluated against every document vector in the database. Relevant documents can be determined by selecting the documents which reside within a close proximity to the query vector. (What documents are most related to the query?)

From the above three evaluations term weight can be calculated with the following

equation:

$$w_i = tf_i * \log (D/df_i)$$

- tf_i is the term frequency at the document level.
- df_i is the frequency of documents that contain term i
- D is the number of documents in the database.
- $\log d/df_i$ is the Inverse Document Frequency.

Greatest term weights belong to terms which have a high frequency at the document level and a low frequency at the database level.

Semantics and the Lexicon

There are innumerable possibilities of search queries which can be passed to a search engine. When word variations like misspellings and stems are considered it becomes apparent that constructing a lexicon of all possible search terms is not possible. Therefore search engines generally employ a two fold system which consists of:

- 1) Root Lexicon – A term database composed of language root words.
- 2) Matching System - Whereby search query concepts are determined and search terms are reduced into their closest representation within the lexicon. One such method is the partial-match system.

Example: Let's say that the following queries are passed to our engine:

Tackle
Attack
Tack

Let's say that our closest Lexicon Term Entry is **TACK**.

Queries would be evaluated as follows:

Tackle: Begins-With Partial Match
Attack: Partial Match
Tack: Exact Match

Obviously the third query term “tack” would receive the highest relevancy score because it is an exact match. The first query term would receive some relevancy because although it is not an exact match, an exact match can be found in the beginning of the word. The second is unlikely to receive any consideration unless no better matches exist, in which instance would only be considered because the target term is found somewhere within the word.

There are some highly advanced language analysis models that are being developed for use by search engines. These systems are generally dynamically arranged vector space models organized according to an intelligent natural language ontology that designed to identify and interpret subjective language usage rules and conventions. The most noteworthy example of such work is the semantic ontology database, SERTA, which was created by Applied Semantics Corporation (Applied Semantics was purchased by Google Inc. in 2003).

Ranking Concepts

Success for a search engine is certainly determined by the engine's ability to retrieve documents that are relevant to a given search query, and successful search engines go to great lengths to protect their specific ranking methodologies. The following definitions will provide a brief explanation of common IR ranking concepts which of course provide the foundation for search engine technology. It is generally agreed that most commercial search engines employ ranking technologies that are based upon these concepts.

Dimensionality Reduction: A Clustering technique whereby document clusters in vector space are considered as a "community" of relevance. The unique feature of this method is that it allows the possibility of selecting documents which contain no instances of a search term (The document being considered relevant because of its association with documents that do contain the search term).

Vector Space Distance: Using the vector space model (as above) a grouping of documents which are closest to the search term vector are collected. Vector distance between the documents and the query are then used to rank the results. Similarity measures such as Jaccard's Coefficient or Cosine Coefficient are used to make this determination. For example:

$\frac{Q \cap D}{Q \cup D}$

QUD = The Intersect of Query weight and Document Weight are Divided by the Union of Query weight and Document weight.

TF/IDF: Counts term frequency in a document which is then normalized by counting the inverse document frequency (global term instances). Assigns a weight to the term itself and a weight to each document for that term.

Citation Text Ranking: The anchor text of a hyper-link is treated as a topical indication of the document to which it points.

Link Graph Inter-Connectivity

Link Graph Concentrations: The internet link graph can be used as an effective indication of the general importance or inherent value of web pages. A web page with many incoming links is considered important and this can be taken into consideration for ranking.

From Google PageRank Technologies, (Green 2004):

“At the First International Conference on the World Wide Web (1994) a researcher named Oliver A. McBryan presented the World Wide Web Worm (WWWW); the first search index which employed document citation as a measure of importance for web pages. Specifically, the WWW used the anchor text of links to documents to determine importance. The Hyperlink Search Engine, developed by IDD Information Services employed back-links as a measure of importance for web pages; when a search was performed, the search term was evaluated against a collection of anchor text descriptions that point to a page, rather than using a keyword index of the page’s content. Hyperlink Search Engine disregarded nearly all page-resident factors when compiling its Search Engine Results Page (SERP). “

The most successful and well known citation based ranking system is Google’s PageRank algorithm, which determines not only the quantity of incoming links but also the quality and thematic relatedness. I have discussed PageRank at great length in my paper: Google PageRank Technologies.

References and Citations

www.miislita.com

Dr E. Garcia

The Anatomy of a Large-Scale Hyper-textual Web Search Engine

Sergey Brin and Lawrence Page

Department of Computer Science

Stanford University

Page Rank Citation Ranking: Bringing Order to the Web

Lawrence Page, Sergey Brin, Rajeew Motwani und Terry Winograd

Department of Computer Science

Stanford University

Efficient Crawling Through URL Ordering

Junghoo Cho, Hector Garcia-Molina, Lawrence Page

Department of Computer Science

Stanford University

United States Patent # 6,725,259

Bharat; Krishna

Google Inc. - April 20, 2004

United States Patent # 6,285,999

Page, Lawrence

The Board of Trustees of the Leland Stanford Junior University - September 4, 2001

www.searchenginewatch.com

Search Engine Watch

www.google.com

Google Inc.

www.yahoo.com

Yahoo Inc.

www.teoma.com

Teoma Search