Map-Reduce Examples

Amit Jain

Associate Professor
Department of Computer Science
College of Engineering
Boise State University

Information Retrieval

- **Term-Frequency and Inverse-Document-Frequency**. The tf-idf weight (*term frequencyinverse document frequency*) is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tfidf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

TF-IDF

- The term frequency (tf) for a given term t_i within a particular document d_j is defined as follows, where $n_{i,j}$ is the number of occurrences of the considered term in the d_j document.

  $$tf_{i,j} = n_{i,j}$$

  The term frequency is often normalized to prevent a bias towards larger documents, as shown below:

  $$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

TF-IDF

- The term frequency (tf) for a given term t_i within a particular document d_j is defined as follows, where $n_{i,j}$ is the number of occurrences of the considered term in the d_j document.

$$tf_{i,j} = n_{i,j}$$

  The term frequency is often normalized to prevent a bias towards larger documents, as shown below:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- The inverse document frequency (idf) is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

  with
  - $|D|$: total number of documents in the collection
  - $|\{d : t_i \in d\}|$: number of documents where the term t_i appears. To avoid divide-by-zero, we can use $1 + |\{d : t_i \in d\}|$.

TF-IDF

- The term frequency (tf) for a given term t_i within a particular document d_j is defined as follows, where $n_{i,j}$ is the number of occurrences of the considered term in the d_j document.

  $$tf_{i,j} = n_{i,j}$$

  The term frequency is often normalized to prevent a bias towards larger documents, as shown below:

  $$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- The inverse document frequency (idf) is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

  $$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

  with

  - $|D|$: total number of documents in the collection
  - $|\{d : t_i \in d\}|$: number of documents where the term t_i appears. To avoid divide-by-zero, we can use $1 + |\{d : t_i \in d\}|$.

- The tf-idf is then defined as:

  $$(tf\text{–}idf)_{i,j} = tf_{i,j} \times idf_i$$

The approach

- **Term-Frequency and Inverse-Document-Frequency**. Given a corpus of text, we want to calculate tf-idf for every document and every token. We need to calculate over the corpus the following: number of tokens, number of unique terms, number of documents, number of occurrences of every term in every document and number of documents containing each term.

Last.fm Hadoop Usage

- Started using in 2006.
- Running on 50 machines, 300 cores and 100TB of disk.
- Hundreds of daily jobs are run performing operations such as logfile analysis, evaluation of A/B tests, ad hoc processing, and charts generation (track statistics).

Last.fm Track Statistics

- **Track Statistics** (simplified version from *Last.fm*). Data comes from two sources.
  - A **scrobble** is a track listen submitted by a user on a music player on PC or mobile device.
  - A **radio listen/play** is a track that a user listens to using Last.fm website or player.
- **Input**: is (user id, track id, scrobble, radio play, skip), where the last three fields are 0 or 1.
- **Output**: Unique number of listeners per track as well as accumulated listens, scrobbles, radio listens and skips.

References

- *Tf-idf*. `http://en.wikipedia.org/wiki/Tf-idf`.
- *Map/Reduce for real problems: Calculating TF-IDF using Hadoop*. explains how to use map-reduce to calculate tf-idf but does not provide any code. `http://romankirillov.info/hadoop.pdf`
- *Hadoop: The Definitive Guide*. Tom White. See Chapter 14 for Last.fm case study.