WebDocs: a real-life huge transactional dataset.

Claudio Lucchese[2], Salvatore Orlando[1], Raffaele Perego[2], Fabrizio Silvestri[2]

[1] Dipartimento di Informatica, Università Ca' Foscari di Venezia, Venezia, Italy,orlando@dsi.unive.it

[2] ISTI-CNR, Consiglio Nazionale delle Ricerche, Pisa, Italy, {r.perego,c.lucchese,f.silvestri}@isti.cnr.it

Characteristics of the dataset

This short note describes the main characteristics of WebDocs, a huge real-life transactional dataset we made publicly available to the Data Mining community through the FIMI repository. We built WebDocs from a spidered collection of web html documents. The whole collection contains about 1.7 millions documents, mainly written in English, and its size is about 5GB.

The transactional dataset was built from the web collection in the following way. All the web documents were preliminarly filtered by removing html tags and the most common words (stopwords), and by applying a stemming algorithm. Then we generated from each document a distinct transaction containing the set of all the distinct terms (items) appearing within the document itself.

The resulting dataset has a size of about $1,48GB$. It contains exactly $1.692.082$ transactions with $5.267.656$ distinct items. The maximal length of a transaction is 71.472. Figure 1 plots the number of frequent itemsets as a function of the support threshold, while Figure 2 shows a bitmap representing the horizontal dataset, where items were sorted by their frequency. Note that to reduce the size of the bitmap, it was obtained by evaluating the number of occurrences of a group of items having subsequent Id's in a subset of subsequent transactions and assigning a level of gray proportional to such count.

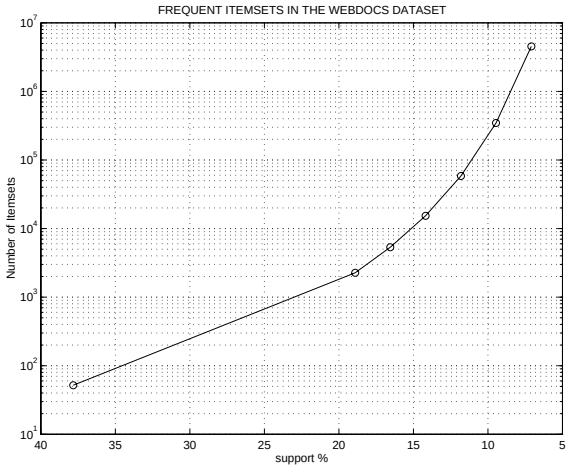

Figure 1. Number of frequent itemsets discovered in the WebDocs dataset as a function of the support threshold.

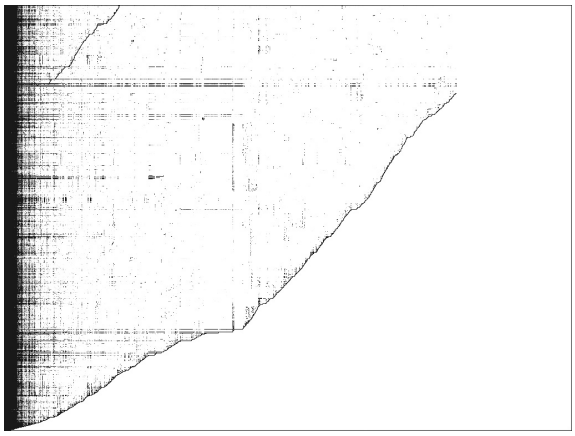

Figure 2. Bitmap representing the dataset.