Web Search Engines — Homework 4

Anirudhan J Rajagopalan
N18824115
ajr619

April 18, 2016

1 Problem 1

1.1 A

Table 5.1 (p. 80) This table shows the effect of preprocessing over various parameters. It shows the subjects in one row and their predicates in rows. So this is a horizontal listing.

Predicate: Filtering techniques
Subject: Columns

Table 5.3 (p. 88) This table is also a horizontal listing.

Subjects: Computer, arachnocentric
Predicate: docIds, Gaps
Structural difference: The difference is that a single predicate is divided into two or more predicate (docids, gaps).

Figure 9.8 (p. 176) This figure is a vertical listing. Vertical listing tables list one or more attributes for a series of similar entities. Here the words are similar entities and the attributes are the nearest neighbours.

Figure 12.3 (p. 221); This table doesnot have any subjects. And the listing is similar to that of vertical listing. Hence by the definition of attribute value listings, this figure is an attribute value listing table.

Structural Difference: The information is contained in the inner attribute value pairs.

Table 14.3 (p. 276) Table 14.3 is an example of Matrix listing. MATRIX tables have the same value type for each cell at the junction of a row and a column. There are two matrices. Here a row is given by training & teseting. And the columns are given by KNN with preprocessing and KNN without preprocessing.

Table 16.1 (p. 323) This is clearly a matrix listing and follows the example above.

Table 16.3(a) (p. 341) This is an attribute value table type.

Attributes: DocID
Value: Text associated with it
Structural diffence: We have four columns instead of single attribute/value pair

1.2 Difficulties of retrieving webtables

Google web tables uses keyword frequency to extract information from tables. It uses multiple algorithms to get the information out of the crawled pages that have tables.

1.2.1 What kinds of queries would be these tables be relevant to? What would be involved in doing that match?

1. These tables will generally be relevant to queries that expect an information from the search engine. Such as what are the docids for 'computer'.
2. There are multiple tasks involved for doing this task. First is to detect the table. Tables can be found in various forms in a HTML. It can be created by table element, or by using clever css or by css grids. Google has to take care of the actual display that is being used for rendering the tables.
3. The next step is to identify the type of tables that are actually displayed. This can fall under one of the several taxonomy.
4. The final step is to extract the actual semantic triplets from the tables and indexing the results.

1.2.2 What kinds of false positives are a danger here? How, if at all, could those be avoided?

There are several dangers while indexing web tables.

1. A table can be assigned to a wrong taxonomy. Can be avoided to a certain extent by having a feedback loop with the user behaviour for the search results.
2. The semantic triplet can be extracted in a wrong fashion. This can be avoided by using a large training corpus to find the actual triplets and also by manually going through a few common pages in the web.
3. Multiple tables and varying triplets. This can be mitigated by identifiying visual clues in the web page.

2 Problem 2