Thesis: Can Jaccard Similarity Index be Applied in Financial Statement Analysis

The purpose of the research is to find out whether Jaccard Similarity Index can be applied in Financial Statement Analysis. The whole research progress currently includes three parts:

Step 1:

Crawl all the zip files from <https://www.sec.gov/dera/data/financial-statement-data-sets.html>, which is a financial statement data set from U.S. Securities AND Exchange Commission. Crawl additional Financial Statement and Exchange Index including Bid price, ask price etc. from <https://www.seekingalpha.com>.

Step 2:

Following the data descriptions in instructions.pdf in the Github Repository, use all the tag.txt and num.txt in every zip file downloaded to conduct a clustering algorithm aimed at determining the major clusters of all the tags.

The following table is a description of TAG.txt, which contains all standard taxonomy tags. Each entry has one ‘tag’, thus one cluster of entries has different ‘tag’s, the approach is to choose the majority ‘tag’ among all the ‘tag’s in each cluster and select the chosen ‘tag’ to represent the ‘tag’ for the cluster. I chose 10 dimensions to represent each entry in TAG.txt, the value in each dimension is a Boolean value. The dimensions chosen should be inclusive and can be represented by the value of other variables in each entry, such as ‘version’, ‘custom’ and something in ‘doc’.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Field Name** | **Field Description** | page7image59913728  **Field Type** | page7image59915648  **Max Size** | **May be NULL** | **Key** |
| **tag** | The unique identifier (name) for a tag in a specific taxonomy release. | ALPHANUMERIC | 256 | No | \* |
| **version** | For a standard tag, an identifier for the taxonomy; otherwise the accession number where the tag was defined. | ALPHANUMERIC | 20 | No | \* |
| **custom** | 1 if tag is custom (version=adsh), 0 if it is standard. *Note: This flag is technically redundant with the version and adsh columns.* | BOOLEAN (1 if true and 0 if false) | 1 | No |  |
| **abstract** | 1 if the tag is not used to represent a numeric fact. | BOOLEAN (1 if true and 0 if false) | 1 | No |  |
| **datatype** | If abstract=1, then NULL, otherwise the data type (e.g., monetary) for the tag. | ALPHANUMERIC | 20 | Yes |  |
| **iord** | If abstract=1, then NULL; otherwise, “I” if the value is a point-in time, or “D” if the value is a duration. | ALPHANUMERIC | 1 | No |  |
| **crdr** | If datatype = monetary, then the tag’s natural accounting balance (debit or credit); if not defined, then NULL. | ALPHANUMERIC (“C” or “D”) | 1 | Yes |  |
| **tlabel** | If a standard tag, then the label text provided by the taxonomy, otherwise the text provided by the filer. A tag which had neither would have a NULL value here. | ALPHANUMERIC | 512 | Yes |  |
| **doc**  page7image41186432page7image59935552 | The detailed definition for the tag (truncated to 2048 characters). If a standard tag, then the text provided by the taxonomy, otherwise the text assigned by the filer. Some tags have neither, and this field is NULL. | ALPHANUMERIC | 2048 | Yes |  |

NUM.txt, SUB.txt and PRE.txt provide the empirical description of each ‘tag’. If the ‘tag’ of a cluster of tags matches a tag in NUM.txt, SUB.txt and PRE.txt, then the empirical financial operation of a cluster of tags is determined.

Next, when a new financial statement report is given, I collect all the Synonyms and Antonyms of the descriptions of 10 dimensions and set the Boolean value on each dimension as 1 if the number of synonyms is larger than the number of antonyms, otherwise 0.

A screenshot of a cell phone

Description automatically generated

I1 [1,1,0,0,0,0,0,0,0,0]

I2 [1,1,1,0,1,1,1,0,0,0]

I3 [0,0,0,1,0,0,0,0,0,0]

I1UI2 ={1,1,1,0,1,1,1,0,0,0}, I1 ∩ I2={1,1,0,0,0,0,0,0,0,0}, then Jaccard Similarity Index is 2/6=1/3, the difference is 1-1/3=2/3. I use hierarchical clustering as the approach because I want to determine the number of clusters after the clustering process.

A screenshot of a cell phone

Description automatically generated

Step 3

After step 2, when a financial statement report is given, the attached empirical financial operation (exchange xxx amount of xxx with xxx) is determined and a back-test function will judge whether the clustering process and the prediction process prove useful——whether Jaccard Similarity Index can be used in financial statement analysis.