

2025 年 春季学期 数据库原理 2 实验 3

2025. 05. 13

Problem Description:

In recent years, detecting similarity among websites has become one of the most prevalent computational tasks, useful for community detection or recommending websites. Nowadays, a wealth of valuable datasets can be found online, which can be leveraged to develop algorithms for similarity detection. For this assignment, you will implement a similarity detection algorithm across three datasets focusing on websites. Here, similarity between websites is measured based on the number of common hyperlinks they share on their pages.

Within a website network, there are two types of relationships. One is symmetric, i.e., website X has embedded a hyperlink to website Y, and website Y also provides a hyperlink directing back to website X; the other is asymmetric, which means website Y may not link to website X even though website X has a hyperlink to website Y. In the latter case, there are two roles involved in this relationship: referrer and referree (akin to relationships seen on Wikipedia). When website X has a hyperlink to Y, X is considered as the referrer and Y is the referree.

The similarity between a given pair of websites is calculated by the number of common referrees divided by the total (deduplicated) number of the two websites' referrees. Specifically, the formal definition of similarity is as follows:

If $|out(A) \cup out(B)| > 0$, we define

$$\text{Similarity}(A, B) = \frac{|out(A) \cap out(B)|}{|out(A) \cup out(B)|}$$

where $out(A)$ is the set of all referrees of website A, and $|A|$ is the cardinality of A. If $|out(A) \cup out(B)| = 0$, we set the similarity to 0.

Example:

The following example demonstrates the process of calculating pairwise similarity among five websites: A, B, C, D, and E. A matrix can be used to illustrate the relationships between these websites. For instance, an element with a value of 1 in the cell corresponding to (A, B) indicates that website A contains a hyperlink to website B. In this case, A is the referrer, and B is the referree. Conversely, an element with a value of 0 signifies that no hyperlink exists from A to B.

According to the matrix above, the set of referrees of A is {B, C, E} and the set of referrees of B is {A, C, E}. There are 2 common referrees between A and B (i.e., C and E), and the number of the union of their referrees is 4 (A, B, C, E). The similarity between A and B is therefore $2/4 = 0.5$.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 1 | 1 | 0 | 1 |
| B | 1 | 0 | 1 | 0 | 1 |
| C | 0 | 0 | 0 | 0 | 1 |
| D | 1 | 0 | 0 | 0 | 0 |
| E | 0 | 0 | 1 | 0 | 0 |

Dataset:

We provide two datasets with different sizes. The small dataset contains around 4K websites and the large one with over 100K websites. Each website is represented by its unique ID number (integer). The small dataset is provided to facilitate your initial debugging and testing.

Sample Input:

The format of the data file is arranged as a referrer-referree pair. As in the above example, it is as follows:

```

A B
A C
A E
B A
B C
...

```

Sample Output: Compute the similarity detection results for EACH website. An example of the output format should be as follows:

```

A B 0.50
A C 0.33
...

```

Objective:

1. Use Flink Table API to analyze the dataset. Print the top 50 lines of your result.
2. Small dataset is used for debugging. If you can get the result from the **medium dataset**, you can get full masks of Homework3.
3. If you can get the result from the large dataset and present how to submit and run the Flink job over Kubernetes in the last week of this semester, you can get full masks of class performance in this course.