Determining the Strength of Relationships between Characters in a story.

Ashwin R. Bharadwaj
Dept. of Computer Science
PES University, Bangalore
PES1201700003
ashwinrb7799@gmail.com

Hardik Gourisaria
Dept. of Computer Science
PES University, Bangalore
PES1201700129
hardik.g@outlook.com

Hrishikesh V.
Dept. of Computer Science
PES University, Bangalore
PES1201700276
hrishi.vish@outlook.com

***Abstract*—The main aim of this project is to develop an algorithm that can identify relationships between entities (Characters of a Novel) and further, use the identified relations to predict the likelihood of the existence of the given entities in a relationship.**

***Index Terms*—Graph, transitive closure, Floyd-Warshall Algorithm, pronoun replacement, pareto distribution.**

I. Introduction

As a proof of concept, the entities are assumed to be characters in a story and the relationship between the characters are the relationships between the entities.

However, the algorithm can find relationships between any entity, living or non-living.

The project is mainly divided into three parts, the first being the identification of entities (characters), establishing the relationship between the said entities and lastly,predicting the possibility of a relationship between entities given the set of known relationships.

The first task of this project was to pick a test data set for proof of concept, for which, Ramayana was chosen. The pre-processing step involved replacing capital letters with lowercase at the start of each sentence. Uppercase letters were only used to identify the characters.

II. Pronoun Replacement

Once the story was chosen, the next crucial step was to replace all the pronouns with proper nouns. We used a greedy heuristic to do so. The assumption was that pronouns usually refer to most recently mentioned character. By that logic, any pronoun can essentially be replaced with the last proper noun encountered while traversing through the paragraph. Care was taken to implement two different stacks for male pronouns and female pronouns.

Proper nouns were pushed to the top of the stack upon encountering them. At the same time, the same character was first removed from the stack to take care of any previous occurrence in the passage. Each character occurs only once in the stack. If a pronoun was encountered, the stack was popped and the first character took the place of the pronoun. This method is accurate about 80% of the time.

A list was maintained, which contained the proper nouns, referring to the characters.

III. Identification of Relationships

Proximity heuristic was used to determine the strengths of the relationships. This means that if two characters are mentioned in the same sentence or in the neighbouring sentences, they are probably related to each other. Greater the number of proximity hits, greater is the strength of the relationship.

To determine all possible relationships between the characters a fully connected symmetric graph(matrix) was used. The values in the graph were initially initialized to zero.

To populate the matrix, the program iterated through the passage (after replacing pronouns with proper nouns) first.The relationship strength value between two characters, A and B, denoted by matrix[A][B], was incremented every time A and B were found in proximity of each other. For Example, if the sentence is ' A and B live together.' then, matrix['A']['B'] += 1.

Once the matrix was initialized with the relationship weights, a few graphs were plotted to view the distribution of the relationships in the story.

From the above graphs, we can infer that the distribution follows a Pareto distribution, which is a skewed, heavy-tailed distribution. The strength of the relationship is inversely proportional to the rank of that particular relationship.

The graph shows that there are a low number of relationships with a high rank and a high number of relationships with lower ranks. It has been shown that any story follows this distribution.

The plot of log of relationship against log of rank is a graph with negative slope, which is consistent.

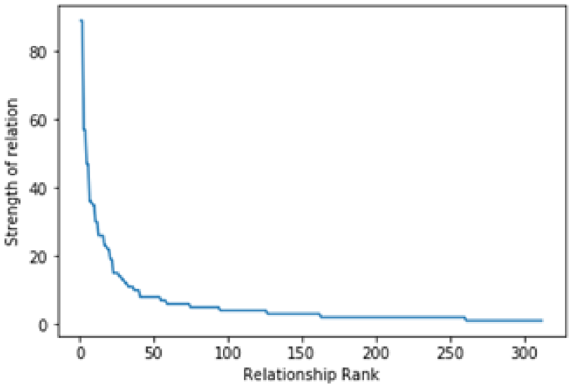

Fig. 1. Relationship vs. Rank

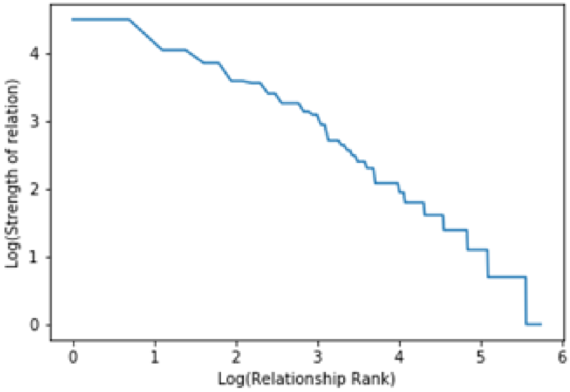

Fig. 2. Log of relationship against log of rank

IV. Conclusion

Usage of graphs and transitive closure to determine the strengths of relationships can not only be used to list the characters in order of importance, but can also be used to conclusively determine if a given character is supposed to exist in the story.

Given three characters, this algorithm checks if the three are meant to be related to each other, if not, it provides an alternate character, who will complete the relationship.