COMP90049 Knowledge Technologies

Project 1 Misspelled Location Names

Bing Xie 741012

2016/9/1

1. Introduction

In computer science, string matching algorithm tries to find patterns within a large string or texts, which is one of effective methods in data mining. In this project, in order to improve the string matching efficiency, dataset problems are discussed. Global Edit Distance and N-gram Similarity are applied and evaluated based on precision with serval value set of threshold. Combine the output data and precision, it can be conclude Global Edit Distance algorithm is better for string matching.

2. Methodology

A program written in python, 'US-loc-names.txt' and 'xieb1_login_tweets_small.txt' of two simplified version dataset are used in this project.

2.1 Datasets Problem Description

2.11 Case sensitivity

Gazetteer records the integer id of geonames and name of geographical point. All geonames begin with capital letter. However, sometimes user might write tweets in all lowercase or uppercase which could cause mismatch and influence the similarity. Given a threshold, if the match is on the boundary condition, case sensitivity could decrease the numbers of matched string, and possibly influence the precision.

2.12 Non-alphabetic character

user_id, tweet_id, and time_stamp in tweet file and geonamesid in gazatter can be regarded as an unnecessary data in matching process. Also, with those unnecessary data, it requires more consuming time to execute the program.

2.13 Duplicate geoname

Gazetteer consists of a large scale of duplicate geonames which can cause long time matching without output if not deal with it. For example, Water well duplicates over 900 times, it may takes over one hour to match and without any matched twitter text. Also, if duplicate geonames have matched tweets text, it will generate redundant matched tweets text. Therefore, it is necessary to remove the duplicate geonames.

2.14 Anomaly twitter content

There exist anomaly metadata in twitter content. Those metadata may not be quite as

obvious. It is discovered when program stop running with warning 'list out of index'. Then I search the last match content 'Scamville' in tweets file, it is an anomy metadata, only a single word 'Scamville' in line, disobey the format of metadata format user_id, tweet_id, tweet_text and time_stamp. After browsing tweets file, there exist numbers of anomy metadata. Therefore it is necessary to take anomy metadata into consideration when program.

2.2 File processor

'US-loc-names.txt' and 'xieb1_login_tweets_small.txt' are the input file for processing.

2.21 Methods for processing Case Sensitivity and Non-alphabetic Characters

According to problem description, in order to effectively match the misspelled location, following steps are applied in tweets file and gazetteer

- Use regular expression to identify all-alphabetic characters except spaces line by line
- Delete all non-alphabetic characters by replacing all matched non-alphabetic characters with space ' 'line by line
- Transform all left alphabetic characters into lower case line by line

By those steps, the user_id, tweet_id, time_stampnd and geonamesid can be removed.

2.22 Methods for merging all duplicated geonames.

Apart from above processing method, another two steps should be applied to gazetteer for merging all duplicated geonames.

- Use data structure set() to automatically identify the location line by line, if a geoname is recorded in a line, the geonames with same name will be identified, and thus all geonames with different names are recorded in set()
- Write each element of set to output processed gazetteer file line by line.

Therefore it requires three steps to process tweets file and 5 steps to process gazetteer.

2.3Methods for solving the approximate string search for misspelled location

Processing the texts into tokens is called tokenization. An easy manipulation is tokens in the query match tokens in the token list of the texts.

2.31 Advantages and Disadvantages of tokenization

Advantages: the list of tokens could be as input for processing such as parsing or text mining.

For this project, tokenization is a necessary step for match the misspelled location since the match strategies require each location as query to match the tokens in tweets file.

Disadvantages: the match precision can be decreased. For example, query is “oneill”, the tokenization for string “O’Neill” could be “o’neill” , “o’” “neill”.

2.32 Global Edit Distance

Global Edit Distance quantifies how dissimilar between two strings.

In this project, four basic operations are used for misspelled location match

- Given location names as query , get the list of strings with same number words of tokens from the tweet file line by line
- Based on query, tokens and global edit distance algorithm, get a list dissimilarity vector.
- Quantify the dissimilarity by

$$overalldissimilarity = \frac{sum\ of\ dissimilarity}{length\ of\ dissimilarity} \times 100$$

- In order to match misspelled location and better performance, threshold is set as 30 and 40. In other word, output match string with similarity larger than 70 and 60.

2.33 N-gram similarity

N-gram similarity quantifies how similar between two strings.

Three basic operations are as following

- Given location names as query , get the list of strings with same number words of tokens from the tweet file line by line
- Based on tokens and n-gram, get a list of n-gram in the string
- Search for the presence of n-gram in query and quantify the similarity by

$$overallsimilarity = \frac{number\ of\ same\ Ngrams}{length\ of\ Ngrams\ in\ query} \times 100$$

- Execute the program with threshold 60 and 70.

3. Results and Analysis

3.1 Evaluation of Global Edit Distance.

Threshold	Match results numbers	Misspelled location	Precision
30	147	13	8.844%
40	1709	34	1.989%

Table 1 Evaluation is conducted with threshold 30 and 40, 1 hour program running time

Query: haverhill public library Similarity:71% Approx. match: charlotte public library Tweet ID: 3194201427 Tweet: charlotte public library not such a bad place to spend a few hours online temp living situation has no internet gasp

Table 2 Output string match

Table 2 shows some of misspelled locations, query 'haverhill public library' match 'charlotte public library' with similarity 71%. Since the recognition of misspelled location manually, the results are as shown in table 1, precision is a fraction of what the system attempt to be correct, and precision can be evaluated as following

$$\text{Precision} = \frac{misspelled\ location\ number}{matched\ result\ number} \times 100\%$$

Then, with threshold 30, system precision is 8.844%. And with threshold 40, system precision is 1.989%. Therefore, for the behavior of system, in terms of threshold, larger threshold will lead to match more misspelled locations and have lower precision.

3.2 Evaluation of N-gram similarity

N-gram	Threshold	Match results numbers	Misspelled location	Precision
3	60	234	3	1.282%
3	70	185	1	0.541%
4	60	105	4	3.81%
4	70	24	1	4.167%

Table 3 Evaluation is conducted on the basis of N-gram 3 and 4, threshold 60 and 70, 2 hour program running time

Query: point b
Similarity:100%
Approx. match: point break
Tweet ID: 3997098911
Tweet: pours one out for point break swayzerip

Table 4 Output string match of N-gram similarity

From Table 3, with threshold rated, larger N-gram will have higher precision. With N-gram rated, the precision have no direct correlation with threshold. Therefore, for N-gram distance, the precision is mainly determined by the magnitude of N-gram.

Table 4 show an example that illustrates the N-gram similarity is not suit for string match since the 100 % similarity corresponds to a location name which has no relation with query, the 'b' is only a substring of 'break'.

3.3 Evaluation on methods for Approximate String Match

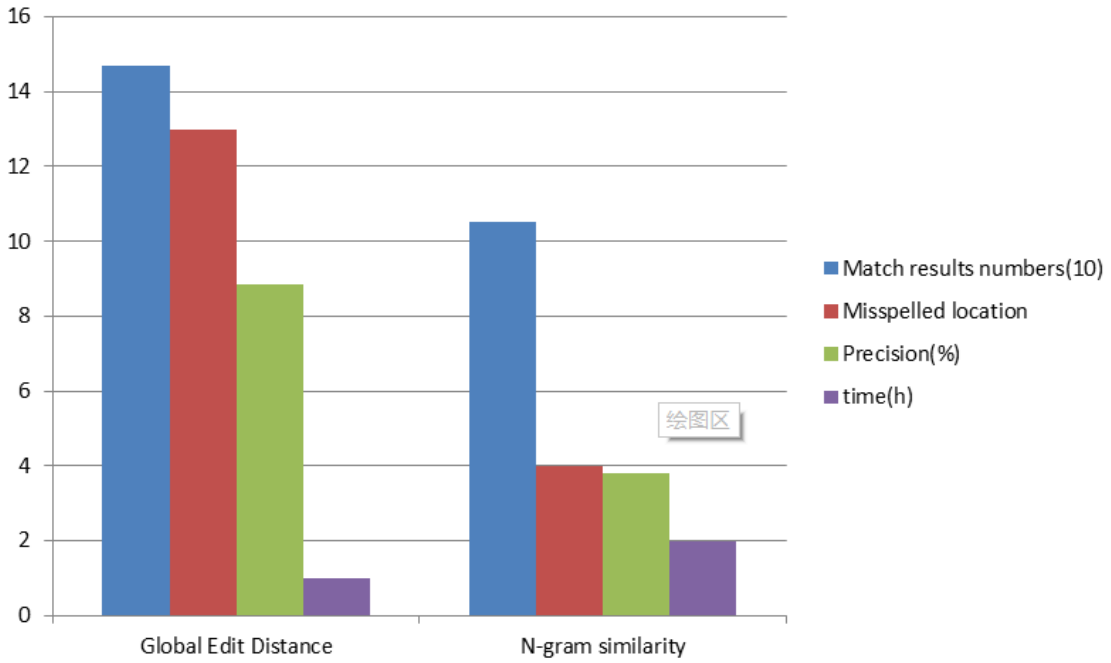

Figure 1 Comparison between N-gram and Global Edit Distance on four aspects

Based on Table 1 and Table 3, Figure 1 can be plot. On the condition that program running time of global edit distance is half of the N-gram similarity, the number of output results (match results and misspelled location) and the deriving result (precision) of Global Edit Distance is larger than N-gram similarity. Therefore, it can be conclude Global Edit Distance match is superior than N-gram for Approximate String Search on four aspects which are

match results number, misspelled location, precision and program consuming time.

4. Conclusion

In conclusion, dataset problems are firstly proposed on Anomaly twitter content, duplicate geoname, non-alphabetic character and case sensitivity. On the basis of solving dataset problems, the file processing methods are discussed for improving the string matching efficiency. Then, Global Edit Distance and N-gram similarity algorithm are implemented on the processed files using different threshold value and different N-gram value in python program. After that, use the precision derived by output data to estimate the Global Edit Distance and N-gram similarity respectively. Finally, comparing the output data, Global Edit Distance is proved to be a better algorithm for string matching on four aspects, which are less time consuming, more output results, more matched results and higher precision.

In future, the program can be optimized to improve the time complexity since the running time is too long to satisfy data mining in future work or research. Better file processing can be researched to decrease the numbers of query. Also, distributed system can be used to compute the output