A Survey on String Similarity Matching Search Techniques

S.Balan1, Dr. P.Ponnuthuramalingam2

1Ph.D. Research Scholar, 2Associate Professor & Head,
Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamilnadu, INDIA.

Abstract: String similarity matching search Problem is mainly used to find text which is present in the documents. In thousands of years many features are available in the modern world but yet people not realized to find the information correctly. Because of huge amount of information’s stored in the World Wide Web. The field of information retrieval was born in the year 1950 and H.P. Luhn in the year of 1957 find the basic idea of searching text with computer. The problem of string matching is to find errors for example in online searching, user faces different problems and irrelevant information’s. The goal of this survey is to present overview of string similarity matching and comparison of different algorithms to conclude the better performance on searching the text. There are many areas where this problem appears and one of the most demanding is information retrieval to find relevant information in text collection and the important tool is named as string matching.

Keywords: Information retrieval, String Matching, Similarity Search, Approximate String Match

I. Introduction

In recent years the problem is growing communities of information retrieval and computational biology. The field of information retrieval problem can be addressed into different views. A string is a sequence of characters over a finite set of alphabet. Similarity search provides a list of input data similar to an input query. In the context of search engines such as Google or yahoo search is based on document similarity and query similarity. Document similarity is nothing but overall similarity of an entire document to the given query. Query similarity suggests many query strings while searching is based on machine learning. [Thomas Bocek, et al., 1997]. At first 1992, text retrieval conference or TREC [Harman 1993] sponsored by US government which aims to encourage research in information retrieval from large text collections.

In that many old techniques are modified and many new techniques are identified to retrieve over large number of text collections. The first algorithms developed in information retrieval for searching the World Wide Web during the year 1996 to 1998. Early there are various models and implementations are available for information retrieval system. Boolean system is used to specify the user information based on combination of And, Or, Not’s. Using this system they are not overcome to produce the relevant information. Several models are proposed for these process in that three most models are vector space model, the probabilistic models, and inference network model [Amit Singhal 2001]. Vector space model is represented by a vector of terms [Gerard Salton, 1975]. Terms are typically words or phrases. Any text can be represented by a vector in high dimensional space. Text belongs to non-zero value. Most vector term processed in a positive value to assign a numeric score to a document for a query. In the year of 1960 maron and kuhn proposed many Probabilistic model and it is based on the general principle that document in a collection should be ranked by decreasing probability of their relevance to a query [Amit Singhal 2001]. Estimation is the key part of this model. Inference network model is a document retrieval model as an inference process in an inference network. [Van Rijsbergen1979] Most techniques implemented under this model. Similarity search is important for time-sensitive applications. The increasing amounts of electronic information available on the web in order to improve data quality or find all information based on the user request. To provide a similarity search in the dictionary size may be too slow for many applications. There are various existing methods are available for fast similarity search for example English dictionary and a randomly generated dictionary and compared search performance for dynamic programming, a keyword tree, neighborhood generation and n-grams with index lookup extraction [Amit Chandel, 2006]. The extraction of structured and unstructured text is a challenging problem in many applications such as data warehousing, web data integration and bio-informatics. For example, to identify book author from html pages, match of text string with book author is displayed and found the accuracy of the string extraction [Amit Chandel, 2006]. This paper categorized into four sections. Section-1 contain the introduction to information retrieval and string similarity search, Section-2 contain the literature survey, Section-3 contain Analysis of string similarity search Section-4 includes conclusion while references mentioned in the last section.

II. Literature Survey

It is defined as a finite state pattern matching machine from the keywords to process the text string in a single pass. To improve the speed of a library bibliographic search program by factor of 5 to 10. The main purpose of
this technique is to allow a bibliographer to find in a citation index all titles and satisfying some Boolean function of keywords and phrases. If \( m \) is a program which takes as input the text string \( s \) and produces as output the locations in \( p \) at which keywords \( y \) appear as substrings. It consists of a set of states and it is represented by a number. The behavior of the pattern matching machine is carried out by three functions named as go to function go, a failure function fa and an output function out [Alfred V. Aho, et al.,1975].

Edit distance [Levenstein V.I, 1966] is the minimum number of operations required to transform one string into another with operations being a deletion, an insertion or a replacement. Navarro’s NR-grep [Navarro.G, 2000] is an exhaustive online similarity search algorithm. NR stands for non-deterministic reverse pattern matching. It uses bit-parallelism and forward and backward searching. An n-gram is created by sliding a window of length \( g \) over the data and noting the content and position of all such windows. An extension of this approach for large text collections uses cosine similarity [Koudas, et al., 2004], \( t \) is a global measure to represent a vector of their frequencies.

Approximate similarity search based on hashing is to hash the points from the database from the probability of higher objects that close to another. It is based on hierarchical tree decomposition for large number of dimensions. There are various algorithms such as locality-sensitive hashing, analysis of locality-sensitive hashing and nearest neighbor search. Approximate string matching is about finding a pattern in a text where one or both of them have suffered some kind of undesirable corruption. The classification and the existing schemes in context of data structure are suffix tree, suffix array, Q-grams, Q-samples. Search approach method is classified into two ways namely partitioning into exact searching and intermediate partitioning based on text and patterns [Kaushik Chakrabartie, et al., 2000].

The existing algorithms are hamming distance, reversals, block distance, Q-gram distance, allowing swaps, approximate searching in multidimensional texts, in graphs, multi pattern approximate matching, non standard algorithms such as approximate or parallel algorithms, indexed searching, these are the other surveys on string similarity matching. There are various string matching types namely multiple string match, extended string matching, regular expression matching and approximate matching. The approximate matching contains various algorithms to find the similarity of given string such as dynamic programming algorithms, computing edit distance, text searching, improving the average case, other algorithm based on automata, bit-parallel algorithms, parallelizing the NFA, parallelizing the DP matrix, algorithm for fast filtering the text, partitioning into k + 1 pieces, approximate BNDM, other filtration algorithms, multi pattern approximate searching, a hashing based algorithm for one error, searching for extended strings and regular expressions.

### III. Analysis of String Similarity Matching Techniques

<table>
<thead>
<tr>
<th>Sno</th>
<th>Author Name</th>
<th>Title</th>
<th>Methods</th>
<th>Advantages</th>
<th>Dis Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alfred V. Aho and Margaret J. Corasick</td>
<td>Efficient String Matching An Aid to Bibliographic Search</td>
<td>Pattern matching algorithm, Construction of go to, output and failure functions, Time complexity of algorithms</td>
<td>Locates keyword in a text string, Directed graph begins at the state 0, Time complexity is large</td>
<td>Substrings may overlap with one another, Partially computed output function</td>
</tr>
<tr>
<td>2</td>
<td>Arvind Arasu, Venkatesh Ganti, et al.;</td>
<td>Efficient Exact-Set Similarity Joins</td>
<td>Threshold based SSJoin, Hamming SSJoin, Jaccard SSJoin</td>
<td>Threshold parameter is high, Vector representation between two sets, Similarity value is 0 or 1</td>
<td>Different similarity sets, Dimension is differ, Common elements</td>
</tr>
<tr>
<td>3</td>
<td>Thomas Bocak, Burkhard Stiller, et al.,</td>
<td>Fast Similarity Search in Large Dictionaries</td>
<td>Edit distance, NR-Grep, N-grams and Cosine Similarity</td>
<td>Minimum operations required from one string to one string to another, Reverse pattern matching, Offline approach</td>
<td>Dictionary size is low, Avoids number of searching words in NR-grep method, Similarity is shared</td>
</tr>
<tr>
<td>4</td>
<td>Kaushik Chakrabarti, Dong Xin, et al.,</td>
<td>An Efficient Filter for Approximate Membership Checking</td>
<td>Pruning condition, Filtering by ISH, Weighted signatures</td>
<td>Three similarity measures are identified, Sub string search is quick, Weighted signature is in decreasing order</td>
<td>Lower bound value is not identified, String similarity is less, Different number of signatures</td>
</tr>
</tbody>
</table>
In this paper, survey focus on various algorithms for string similarity matching based on search techniques. Some of the algorithm for set similarity with its property value is 0 or 1. It indicates the previous algorithms matches more than in many cases. The performance of the algorithm is analyzed and stated in a table manner. Additionally it focuses on information retrieval and search engine in World Wide Web. To improve the quality of a word search similarity, next the exact similarity is finer based on semantic relationship of a word. This further reduces the time size for a large database.

IV. Conclusion

In this paper, survey focus on various algorithms for string similarity matching based on search techniques. Some of the algorithm for set similarity with its property value is 0 or 1. It indicates the previous algorithms matches more than in many cases. The performance of the algorithm is analyzed and stated in a table manner. Additionally it focuses on information retrieval and search engine in World Wide Web. To improve the quality of a word search similarity, next the exact similarity is finer based on semantic relationship of a word. This further reduces the time size for a large database.

V. References

[13] Thomas Bocek, Burkhard Stiller, et al., Fast Similarity Search in Large Dictionaries, University of Zurich, Department of Informatics (IFI), Binningulustrasse 14, CH-8050 Zürich, Switzerland, 2007.