ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 10, October 2016

An Efficient Approximate Search Using String Transformation; A Survey

Kishor Mahale[1] ,Prof. Miss. Khusbhu Sawant[2] ,Prof. Kuntal Barua[3]
PG Scholar, Dept.Of Computer Science & Engg., JDCT,Indore, M.P., India[1]
Professor, Dept.Of Computer Science & Engg., JDCT,Indore, M.P., India[2]
HOD, Dept.Of Computer Science & Engg., JDCT,Indore, M.P., India[3]

ABSTRACT- String transformation can be considered as a problem in natural language processing such as data mining, information retrieval, bioinformatics, medical science etc. Generally there is a need to change the input string in data mining, natural language processing, information retrieval and bioinformatics. Many times the user is not from the technical background so he may enter the incorrect input string. Most of the time for better results of the search, there is need to transform the input string. Sometimes the user also enters the short forms that are abbreviations for the search; in that case there is need of transforming these abbreviations into their original forms or meanings. Thus there is need of converting abbreviations into their original form, correction of spelling errors and also replacing the word with its synonym if needed for better search results. Thus these conversions of strings can be stated as string transformation. If we simply consider the medical system there is great need of string transformations in the system. There are systems which uses different methods of string transformation and generation for giving better results. String transformation can be conducted in two different ways depending upon the use of dictionary that is whether the dictionary is used or not. In this approach log linear model is expressed in terms of an input and output strings. The method uses an approach for string transformation which is both accurate and efficient. Thus different string transformation methods are used and queries are reformulated for getting accurate as well as efficient results. An algorithm is used to find the top K matching candidates. According to experimental results on large scale datasets the proposed method is accurate and efficient on different string transformation methods.

KEYWORDS- String Transformation (ST), Spelling Error Correction, Query Reformulation, Top K pruning, Log Linear Model.

I. INTRODUCTION

There has been great research and development in the field of data mining, natural processing of language, bio-informatics, medical sciences etc. For getting the information through the large database different search systems and technologies have been developed. There are many algorithms designed to get the accurate search results. Even though there are very good systems, it has been observed that it depends on the user to get the accurate results. It means the system gives the accurate results only if the user enters the perfect or right query. So to get accurate results the system should fix with perfect query. The e orts which are made in developing the search engines become less effective if user does not enter the perfect or right query. Investigations say that not only the non-technical users but the technical users also make mistakes in the query while searching.

It has been observed that many researchers have proposed and developed different technologies for better string search. There are also different methods proposed for string transformation for effective search. So why not to use these technologies for accurate and efficient search. It has been observed that in the medical field the drugs, brand names are too di-cult to remember their spellings and remember it. Also the dataset is too large and string length is also more. So for helping the medical person to search these drugs/brands a system can be implemented. Here both

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 10, October 2016

purposes will be fulfilled string transformation for efficient and accurate search. The technologies which have been developed are mostly based on web search. But here a custom database is developed on which the system is implemented. The main purpose of the dissertation is achieved on this datasets. If required we can also connect the system to web and obtain the efficient results. Suppose the user wants more details about the entered query from internet then this facility is also provided in the system.

II. LITERATURE REVIEW

1. String Transformation

Generation of one string from another can be considered as string transformation. For example we can generate three different meanings of HCL as "Hindustan Computer Limited" or "Hindustan Copper Limited" or "Hydro Chloric". So depending upon the query the short forms HCL will be replaced by the appropriate full string. Similarly if we consider the medical database there are many short forms used for different strings. For example BP in medical terms stands for blood pressure. Thus for the exact and accurate search there is need to replace these short forms by their original meaning. Many researches are made for string transformation such as Arasu el.al proposed a method which focuses on the coverage of rule sets. Tejada el.al proposed an method which estimates the weights of transformation rule with small user input[3]. Okazaki incorporated the predefined rules such as stemming, pre x, suffix, acronym in L1-regularized logistic regression model and utilized it for string regeneration [6].

2. Approximate String Search

In medical terms there are many strings which are almost similar to each other. There is just a difference of one or two alphabets in this strings.so depending on the query the exact string can be selected. The approximate string can be found by two methods 1) using dictionary and 2) without using dictionary. It is assumed that here the string will be chosen with the help of dictionary only i.e. dataset. It is assumed that in approximate string search the model is fixed and the objective is to efficiently find all the strings in dictionary the existing methods uses N-gram based algorithms or tries based algorithm for finding candidates with a fixed range. There are also methods which uses n-grams for finding the top k candidates [5].

3. Spelling Error Corrections

Spelling mistake correction generally contains generation of candidates and selection of candidates. Generation of candidates is mostly related to a single word. Suppose we are having a single word, some rules are applied for spelling correction [14]. The edit distance method is used which typically performs deletion, insertion and substitution of characters. Some methods uses x range of edit distance while some uses different ranges. Edit distance is not concerned with the misspell characters. Some researchers have been done for context based words. A generative model has been developed by Brill and Moore [10] which includes the contextual substitution rules. Further this model was improved by adding pronunciation factors by Toutanova and Moore[8].In my approach the user will be provided with different k output strings for suggesting the spelling correction depending upon their ranks which are searched mostly and most suitable matching word.

III. PROBLEM DEFINITION

Many problems in natural language processing, data mining, information retrieval, and bioinformatics can be formalized as string transformation. In data mining for effective search there is need to change the string so that the system produces accurate results. There are different search engines designed and developed for effective search.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 10, October 2016

Some of them are generative model, logistic regression model and discriminative model. The main focus of these models is accuracy. As the researches and developments are growing along with that data is also growing fast. Thus in today's world users not only demand for accuracy but also for efficiency. Thus the search engine should give accurate results in less possible time even if the dataset is large.

Although the main aim of the system should be to find out the more detailed data from the search system. Like if the user enters the symptoms then the disease name, treatment details should be displayed and also some cases of the patients. But considering the time limit and availability of the database only the drugs basic information which is provided by the chemist/pharmacist is been displayed in the system. In this system it is tried to get most accurate search results in less time. The system should not confuse the user by giving the irrelevant results. Thus the ranking is used to get the most likely searched strings so that the irrelevant data won't be displayed.

IV. PROPOSED SOLUTION

To take care of the above expressed issue distinctive approach is required. A novel and probabilistic approach is connected to take care of the above expressed issue. There are two stages utilized as a part of the framework learning and era stage. The learning stage incorporates the information and conceivable yield strings in the wake of applying distinctive standards. What's more, the era stage incorporates era of various string contingent on the client entered look string. A medicinal medications database is taken as the word reference database. The log direct model is utilized to get the info yield strings. The top-K pruning calculation is utilized to locate the most reasonable yield strings. Likewise for more productive and precise results positioning is utilized so that the client will get just highest yield strings.

Figure 1 indicates framework design of general framework. This framework fills in as takes after: Client wills first change over the piece information in RDF information design i.e in triplet arranges. It will be kept as a database and utilized via web search tool for seeking watchwords. At whatever point User will enter a catchphrase it will look in web crawler and will seek in dataset. Also, last result will showed to the client. Principally there are two sections of this framework. Learning stage and era stage. Learning stage is at a foundation level. Firstly the information is considered i.e database is examined and required string sets are discovered. Contingent upon the string sets, diverse standards are separated which can be connected for string change. Accordingly the principles are examined and a model is created which is utilized for the di erent string change. The second stage incorporates the real execution of the framework. At the point when the client enters the string the framework begins recommending the conceivable related string recommendations. After the client enters they inquiry, the question is circulated in parts into various strings. At that point these strings are checked in the word reference, if not discovered then framework begins applying distinctive principles on the strings so that based yield string can be chosen. In this manner in the wake of selecting the best reasonable yield string the inquiry is reformulated and goes for further preparing. Therefore the framework executes the red question and shows the yield to the client. Taking after are diverse strides in question reformulation.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 10, October 2016

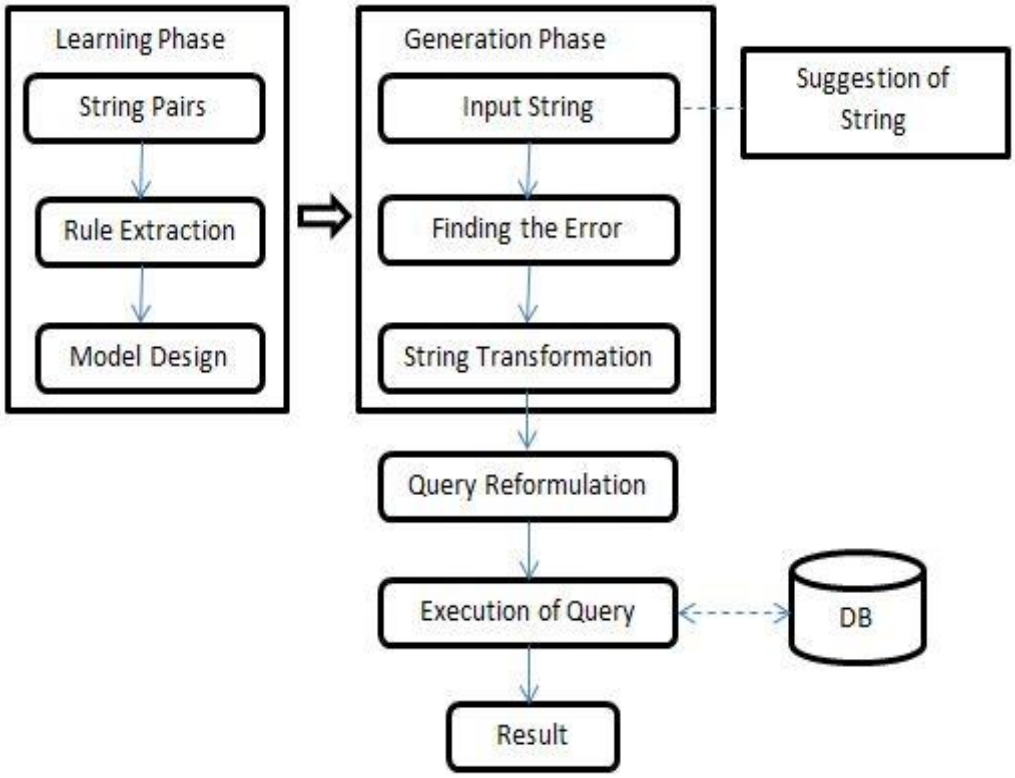

Fig. 1 System Architecture

V. EXPECTED RESULTS

- To get best search results using the string transformations.
- The search results should not be approximate but it should be accurate.
- Along with accuracy main objective is achieving efficiency.
- System should work same even if the data size increases.
- The main objective is to enhancing both accuracy and efficiency in search.
- Applying different rules for string transformation.
- To find the top k output strings by applying different operators.
- Query reformulation by using string transformation.
- To check the spelling errors and get the exact string from approximate string

VI. CONCLUSION

In this system user is provided with extra facilities of search. User is allowed to enter any string for search. Mainly if the user is from non-technical background and has very few details of spelling of strings then it's easy to use this system. Another feature of this system is that search results are very accurate. Also the main concentration is not only on accuracy but also on efficiency. This is an important feature of this system even if the datasets increases the efficiency does not degrade.

The main aim was to achieve efficiency along with the accuracy. The experimental results show that even if the dataset increases the search time does not increases much.

The search results are found in effective time only. This system is especially helpful for the medical science people where there are more chances of spelling mistakes. Also especially designed for the non-technical user who can

probably make spelling mistakes.

REFERENCES

[1] Ziqi Wang, Gu Xu, Hang Li, and Ming Zhang "A Probabilistic Approach to String Transformation" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:PP NO:99 YEAR 2013.

[2] M. Li, Y. Zhang, M. Zhu, and M. Zhou, "Exploring distributional similarity based models for query spelling correction," in Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ser. ACL' 06. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 1025-1032.

[3] A.R. Golding and D. Roth, "A winnow-based approach to context-sensitive spelling correction" Mach. Learn, vol. 34, pp. 107-130, February 1999.

[4] J. Guo, G. Xu, H. Li, and X. Cheng, "A unified and discriminative model for query refinement" in Proceedings of the 31st annual international ACM SIGIRconference on Research and development in information retrieval ,ser. SIGIR'08, New York, NY,USA: ACM, 2008, pp. 379-386.

[5] A.Behm, S. Ji, C. Li, and J. Lu, "Space-constrained gram-based indexing for e-cient approximate string search" in Proceedings of the 2009 IEEE International Conference on Data Engineering, ser. ICDE'09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 604-615.

[6] E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction" in Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ser. ACL '00. Morristown,NJ,USA: Association for Computational Linguistics,2000, pp. 286-293.

[7] N. Okazaki, Y. Tsuruoka, S. Ananiadou, and J. Tsujii, "A discriminative candidate generator for string transformations" in Proceedings of the Conference on Empirical Methods in Natural Language Processing,ser. EMNLP '08, Morristown,NJ,USA: Association for Computational Linguistics,2008,pp. 447-456.

[8] M. Dreyer, J. R. Smith, and J. Eisner, "Latent-variable modeling of string transductions with nite-state methods" in Proceedings of the Conference on Empirical Methods in Natural Language Processing,ser. EMNLP'08. Stroudsburg, PA, USA: Association for Computational Linguistics,2008,pp. 1080-1089.

[9] A.Arasu, S. Chaudhuri, and R. Kaushik, "Learning string transformations from examples" Proc. VLDB Endow.,vol. 2,pp. 514-525,August 2009.

[10] S. Tejada,C. A. Knoblock, and S. Minton, "Learning domain independent string transformation weights for high accuracy object identi cation" in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD'02. New York, NY, USA: ACM,2002,pp.350-359.

[11] M. Hadjieleftheriou and C. Li, " E-cient approximate search on string collections" Proc.VLDB Endow.,vol.2,pp.1660â□“1661,August 2009.

[12] C. Li, B. Wang, and X. Yang, "Vgram: improving performance of approximate queries on string collections using variable-length grams" in Proceedings of the 33rd international conference on Very large data bases,ser. VLDB'07. VLDB Endowment, 2007,pp.303-314.

[13] X. Yang, B. Wang, and C. Li, "Cost-based variable-length-gram selection for string collections to support approximate queries efficiently" in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ser. SIGMOD'08. New York,NY,USA: ACM,2008,pp.353-364.