ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Statistical Learning Method for String Transformation

Sneha Mary Thomas, Nimmymol Manuel

Student, Dept. of Computer Science and Engineering, Mangalam College of Engineering, M.G University, Kerala, India

Assistant Professor, Dept. of Computer Science and Engineering, Mangalam College of Engineering, M.G University, Kerala, India

ABSTRACT: The string transformation is the process in which the system can generate output strings which are closer to each other with respect to the given input string. The string transformation process can be mainly applied for various problems in natural language processing, data mining for database record matching. The string transformation can be used for various applications. The process of string transformation should be conducted in an efficient manner and with higher accuracy for generating the output strings. In most of the existing methods of string transformation efficiency and accuracy is not much considered into account. The proposed method uses statistical learning method for string transformation in which string and query generation is being performed. The query generation process uses three queries which selects the data's from the database and an object or record based method is proposed. The object based method is much efficient and uses less memory when compared to the other three methods. The proposed method is much efficient and accurate method in which semantic meaning retrieval, error corrections in spelling and reformulation of queries is being performed. The experimental result shows that the proposed method for string transformation using the statistical learning method is much accurate and efficient method.

KEYWORDS: string transformation; semantic meaning retrieval; object based method; reformulation of queries

I. INTRODUCTION

The process in which knowledge is being discovered from huge amount of data is known as the data mining process. The string transformation process can be applied for various applications of data mining. Inconsistencies and noises may occur in data and through the process of data cleaning the data become consistent that is noises are removed from the data to make it cleaner. The process in which the system can generate output strings which are closer to each other with respect to the given input string is referred to as string transformation process. The string transformation process can be used in error corrections in spelling and reformulation of the queries. String transformation is the process in which the system generates output strings which are closer to the input string which is being provided. The string transformation process is shown in fig.1.

The string transformation can be conducted by using a dictionary and also without using the dictionary. If the string transformation process is carried out using a dictionary then the output strings produced by the string transformation should be present in the dictionary. In most of the existing methods of string transformation efficiency and accuracy is not much considered as an important factor. The proposed method uses statistical learning method for string transformation in which string generation and query generation is being performed. The string generation process mainly generates all combinations of the input string as output strings. In the query generation process three queries are being used which selects the data's from the database and an object or record based method is proposed. The gram based method, neighbour based method and incremental methods are the three queries which are being used for retrieving data's from the database. The object or record based method is much efficient method and uses less memory when compared to the other three methods. The proposed method is much efficient and accurate method in which semantic meaning retrieval, error corrections in spelling and reformulation of queries is being performed.

DOI: 10.15680/IJIRCCE.2015. 0308057

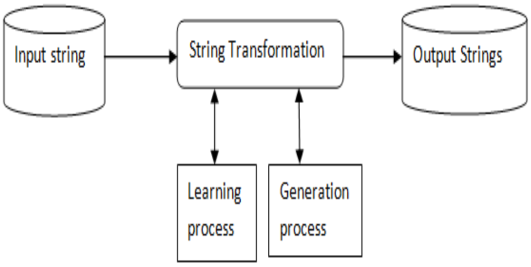

Fig. 1 . The String transformation process

The string transformation process uses the learning process and generation process for generating the output strings from the given input string. The authors in [1] proposed a probabilistic method for string transformation which uses a log linear model for the string transformation which is accurate for spelling error correction and query reformulation. The search as you type method introduced by the authors in [2] uses three queries for retrieving data's from the database.

II. Related work

The string transformation process can be used to formulize various problems in natural language processing, data mining, etc….Most of the methods for string transformation process is less efficient and less accurate. In [1] authors proposed a probabilistic method for string transformation which uses a log linear model. The log linear model is a method used for training the model and also it uses an algorithm for generating the output strings in the string transformation process. The probabilistic method mainly consists of two processes for string transformation namely the learning process and the generation process. The learning process consists of training string pairs from which the rules will be extracted. The learning system is used to construct the model for string transformation. The generation system in the generation process generates the output strings corresponding to the new input string given by referring to the model. For improving the web search the query speller important for the search engines. In [4] authors proposed a method that uses the distributional similarity obtained from query logs for query spelling corrections. The method uses the distributional similarity occurring between two terms and the distributional similarity between two terms will be high between a frequently occurring misspelling and its correction the distributional similarity between two terms will be low between two irrelevant terms with similar spellings. Machine learning problems in natural language mainly uses the characterization of linguistic context. In [5] authors proposed a winnow based method for context sensitive spelling correction. The context sensitive spelling correction is the process of finding out errors occurring in the spelling. The context sensitive spelling correction involves the characterization of linguistic contexts where different words occurs. The winspell algorithm is being used which improves the performance of the method. The correction of spelling errors in the query mainly consists of two methods. The methods are candidate generation and candidate selection. The most likely corrections of the misspelled word is being found out using the candidate generation process. The candidate generation process is being considered with a single word and after the candidate generation process the words in the query can be used for the final candidate selection process. The process of object identification occurs when integration of data's takes place from various websites. The active atlas system proposed by authors in [6] uses domain independent string transformation method for the object identification process. The domain independent string transformation is used for object identification where similarity score is being used to check whether the objects occurring at various sites are similar. The various applications including spelling error corrections use the improved error method. In [7] authors proposed a method for the noisy channel spelling correction. The improved error method mainly consists of two components. The first component is a source model and the second component is a channel model. The model for spelling correction is based on generic string to string edits. In [8] authors proposed an interactive fuzzy keyword search in which the system searches the data as soon as when the user types in the query keyword. The method is an interactive method since the user who has limited information about the data also can retrieve data's using the fuzzy keyword search.

III. PROPOSED METHOD.

The string generation process mainly generates all combinations of the input string as output strings. In the query generation process three queries [2] are being used which selects the data's from the database and an object or record based method is proposed. The gram based method, neighbor based method and incremental methods are the three queries which are being used for retrieving data's from the database. The object or record based method is much efficient method and uses less memory when compared to the other three methods. The proposed method is much efficient and accurate method in which semantic meaning retrieval, error corrections in spelling and reformulation of queries is being performed. In the incremental method for data retrieval the data's will be retrieved by removing each character from the input string which is being provided.

The proposed method provides an object or record based method which selects the data's from the database. The object based method uses less memory and time for data retrieval when compared to the other three methods. The proposed method supports error corrections in spelling and semantic meaning retrieval. In the process of error corrections in spelling the errors occurring in spellings can be corrected and in the process of semantic meaning retrieval the semantic word corresponding to the input string will be retrieved from the database.

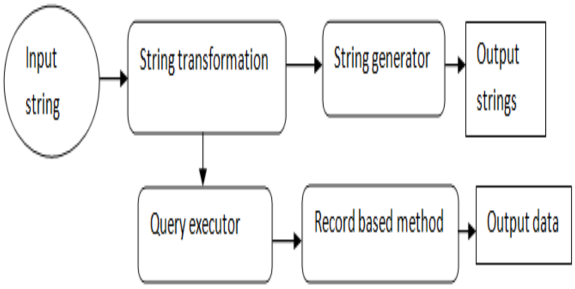

Fig. 2 . The Proposed Method

The main working of the proposed method is shown in fig.2. The proposed method consists of string generator and query executor. The string generator generates the output strings from the given input string and the query executor performs the queries to retrieve data from the database and the record based method is being performed with the queries which improve the performance of the queries.

IV. EXPERIMENTAL EVALUATION

The fig.3 shows the performance of the gram based method when compared with the object or the record based method. The gram based method uses much higher memory when compared to the gram (record) based method. When the gram based method alone is performed the memory usage will be higher and when the gram based query is performed using the record based method the memory usage will be very less. The graph shows that for the gram(record) based method as the database size increases the memory usage decreases. The proposed method improves the accuracy of string transformation process. The fig.4 shows the comparison of the proposed method and the existing method. The proposed method provides much higher accuracy when compared to the existing methods. As the number of strings increases the accuracy for the proposed method also increases. The proposed system provides a higher efficiency and accuracy for the string transformation when compared to the existing system.

DOI: 10.15680/IJIRCCE.2015. 0308057

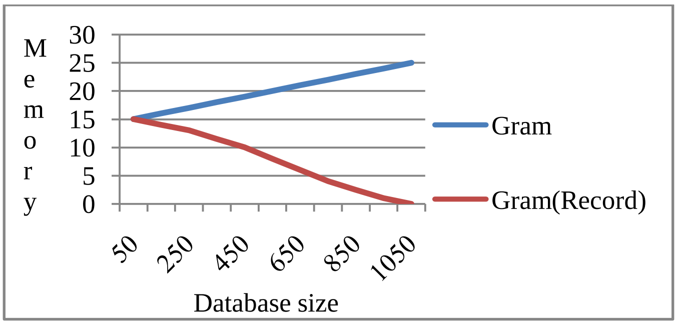

Fig. 3 . Performance of gram based and record based method

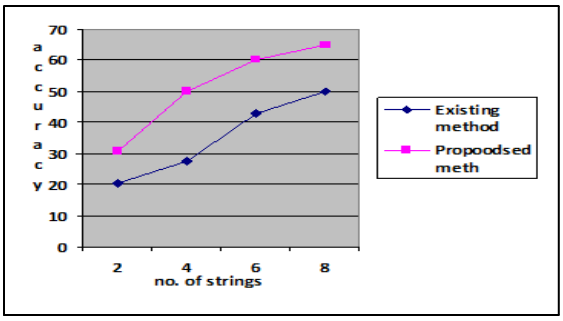

Fig. 4 . Comparison of proposed method with existing method

V. Conclusion and Future Work

The string transformation process is the process in which the system can generate output strings which are closer to each other with respect to the given input string. The proposed system uses string transformation for error corrections in spelling and semantic meaning retrieval. The proposed system uses the record based method for data retrieval in the query generation process. The experimental results show that the proposed system provides a higher accuracy when compared to the existing system. In future the string transformation can be applied to more applications.

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. 26, No. 5, pp. 1063-1075, May 2014.
2. Guoliang Li, Jianhua Feng and Chen Li, 'Supporting Search-As-You-Type using SQL in Databases', IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 2, pp. 461-475, February 2013.
3. M. Hadjieleftheriou and C. Li, 'Efficient approximate search on string collections', Proc. VLDB Endow., vol. 2, no. 2, pp. 1660–1661, Aug. 2009.
4. M. Li, Y. Zhang, M. Zhu, and M. Zhou, 'Exploring distributional similarity based models for query spelling correction', in Proc. 21st Int. Conf. Computational Linguistics and the 44th Annu. Meeting Association for Computational Linguistics, Morristown, NJ, USA, pp. 1025–1032, 2006.
5. A. R. Golding and D. Roth, 'A winnow-based approach to context-sensitive spelling correction', Mach. Learn., vol. 34, no. 1–3, pp. 107–130, Feb. 1999.
6. S. Tejada, C. A. Knoblock, and S. Minton, 'Learning domain independentstring transformation weights for high accuracy object identification', in Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, New York, NY, USA, pp. 350–359, 2002.
7. E. Brill and R. C. Moore, 'An improved error model for noisy channel spelling correction', in Proc. 38th Annual Meeting Association for Computational Linguistics, Morristown, NJ, USA, pp. 286–293, 2000.
8. S. Ji, G. Li, C. Li, and J. Feng, 'Efficient interactive fuzzy keywordsearch', in Proc. 18th Int. Conf. World Wide Web, New York, NY, USA, pp. 371–380, 2009.

BIOGRAPHY

Sneha Mary Thomas is a student doing M.Tech in the Computer Science And Engineering Department, Mangalam College of Engineering, M.G University, Kerala, India.

Nimmymol Manuel is an Assistant Professor in Computer Science And Engineering Department, Mangalam College of Engineering, M.G University, Kerala, India.