A Heuristic Approach to Enforce String Transformation using Ontology and Log Module

Ketaki Ganesh Katre
PG Student, Department of Computer Engineering
JSPM's ICOER,Wagholi
Pune, India

ABSTRACT
Searching proper information is becoming most challenging task due to increasing amount of information in the web. Search engines smartly do this thing to fulfill the user's requirement. But Search engines are packed with billions of URL's and there are millions of permutation and combination of the keywords to provide the query for search engines. So, to ease this process of firing query where user can come to know about the query string as he is entering some consecutive characters for the query. And this is known as the String transformation technique. Many methods are been introduced to provide service for this technique, but most of them are not relay on the meaning of the String. So proposed system put forwards an idea where semantic of the word is identified using the ontology. Using generalized inverted index actually speed-up the process of searching.

General Terms
Algorithm

Keywords
String transformation, Ontology, Generalized inverted index, Log linear model, Protégé Tool.

1. INTRODUCTION
String transformation is one of the basic method used to solve the numerous problems in the field of information retrieval, NLP, data mining etc. when we type any word or character on web browser then the words relevant to that character will get displayed.

There are number of problems that come under the string transformation category like spelling correction, stemming, and pronunciation generation etc. String transformation method can be used for the suggestion of query or reformulation of query while doing searching. Also in data mining it is used for the synonyms matching and matching of database records. To do string transformation two different settings can be used one is to use dictionary or another is without dictionary.

When the dictionary is used then the output words are searched from the dictionary however size of the dictionary can be very large. To utilize the dictionary with great efficiency indexing is done to each string. Such that dictionary must have path from the root node to the leave node. When we dealing with substring at that time we expands the path , then we check whether the expansion from the root are valid or not if the expansion is not valid then we discards it. Consider an example, let current path represents a string mic. Now there are two ways by which we can make expansion. First is to use transformation rule o->u and another is o->ro. But in dictionary "c" does not have "u" as child node so this expansion will not take place. In this way selection of path is done.

So to overcome this drawback numbers of approaches are proposed which makes use of substitution rules that contains the contextual information of the words. E. Brill et.al. Explains a method in which spell correction model is generated which depends on the context based substitution rules [1]. Mine spelling-error and correction pairs is a pair that focuses on the human behaviour of misspelling words while browsing web. This pair can be effectively used in correction of spellings.

String transformation can be used to correct the spelling errors. Normally it consists of two steps i.e. candidate selection and candidate generation. However candidate generation is used only for the single word. Commonly for the candidate generation of single word a rule-based approach is used. Edit distance is a preferred approach which consists of character insertion, deletion and substitution. There are some methods which makes use of edit distance of fixed range. Edit distances will not takes context into consideration. For e.g. most of the time people's misspelled "c" as "k" or "s" if context is considered but edit distance is not capable to deal with such situation.

Huizhong Duan et.al. Developed a spelling correction model that can correct the spellings online [2]. This method makes use of Markov n-gram transformation model to correct the spellings. In this paper different issues related with spelling corrections are addressed.

Query reformulation is an important method of searching process. This method rewrites the original words by substitution words to increase the efficiency of the searching techniques. Normally it is used to deal with the mismatch issue of the word. Consider query "USA" and the documents contain "united states of America", and then such documents will not get ranked because the query will not match with the content of the documents. Thus query reformulation is done that converts the word"USA" to the "United states of America". By doing so effective matching of words can be done and obviously this will increase the probability of ranking.

However string transformation model is categorised in to two groups. First group assume that the model is given to generate the efficient strings. On other hand model includes different approaches such as generative, discriminative, regression etc. In second approach efficiency is not considers as a important factor. While defining string transformation methods three problems get rise up. 1. How to generate system to achieve both accuracy and efficiency, 2. How to derive the instances from the training instances, 3. How to generate top k strings based on the input query.

Rishabh Singh et.al . Elaborates a method that gives a semantic string transformation [3]. Here the string transformations are well explained by the examples. Here

spread sheet table is considered which is small in size and due to this it increases the performance of the system.

The rest of the paper is organized as follows. Section II discusses some related work and section III presents the design of our approach. The details of the results and some discussions we have conducted on this approach are presented in section IV as Results and Discussions. Sections V provides conclusion and future scope.

2. RELATED WORK

In today's world of data explosion, large information is needed to review to extract the particular things. However the extraction scheme used should be scalable enough so that the performance of the system will not degrade. To extract the particular documents from the large collection of data a keyword searching technique is used.

There are number of approaches has been proposed to search keywords from the large datasets. The algorithm like page rank which used by the Google, HITS are some of the well-known searching algorithms. In keyword searching algorithms a keyword is feed as a input to the algorithm and from that keyword a relevant documents such as pdf, text, xml will be return by the algorithm. Apart from all this a data structure known as Inverted index is one of the best solutions to search the particular word from the huge collection of data.

Almost all the searching techniques that need to be search on textual data make use of inverted index. In retrieval techniques the things to be retrieved are known as documents irrespective of their nature like pdf, doc, web etc. In inverted index technique from a given keyword a list of document is generated, this list contains all the documents that contain the query words.

Consider the example

Table 1(a): A sample dataset of 8 file contents

ID	Content
1	Online service is popular in pune
2	Pune is good place for online bank
3	Many people use online bank services in pune
4	Online service is faster
5	Bank provide service for customer
6	SBI is most preferable bank in pune
7	Huge amount of online service available in pune
8	Pune is place of online bank

Table 1 (b): Inverted index

Word	IDs
Online	<1,2,3,4,7,8>
Bank	<2,3,5,6,8>
Pune	<1,2,3,6,7,8>

In this way inverted index is generated. It contains all the ids of the documents in which query word is present.

Ontology is one of the best approaches used in information retrieval techniques for efficient retrieval data. Ontology is a hierarchical structure used to represents an entities having interdependent relationship with each other. The concept of ontology is widely used in information retrieval, indexing techniques, artificial intelligence areas etc. In text retrieval techniques, Ontologies are used to fill the gap between the texts based web pages and semantic data on the web. One of the main disadvantage of the inverted indexes is they ignored the geographical reference. Consider example hotels in India, Here India will treat as simple word only and hence only those documents containing India word will get retrieved. Here the semantic of the India will not take into consideration. Hence such problems can be easily overcome by using ontology.

Miguel R. Luaces et.al. Illustrates a new concept in which three different techniques such as inverted index, spatial index and ontology based structure is used to retrieve the documents [4]. Here ontologies are combined with the indexes to capture both i.e. statistical and geographical relationships between the documents. In this methodology ontology is created to keep the geographical references of the words.

Sajendra Kumar et.al. Gives another study which makes use of ontologies to retrieve the documents [5]. Here new semantic based index structure is proposed. This paper makes use of ontology and natural language processing to generate the index structure. Therefore context based retrieval of documents is done. Thanh Tran et.al. Proposed a similar approach where keyword queries are mapped to the DL conjunctive queries [6]. To do so it makes use of KB, from the information presents in KB mapping is done.

Whenever user browse the internet to search particular things, it may possible that he spell the word in wrong manner. During such scenario a wrong spelled word should get replaced by the correct spelling. A numerous approaches are proposed to do so. Method that generates correct spelling from the wrong one is known as string transformation method. In spelling error correction method a dictionary is used which contains the proper words. A dictionary is well indexed so the operations will get speed.

Jeyalakshmi.S et.al. Gives an outstanding approach for query transformation [7]. In this paper problem information extraction due to query is addresses. Here commentz Walter algorithm is used. This algorithm is used to achieve more accuracy. Experimental evolution on two large datasets shows that the method is efficient for query transformation in terms of accuracy and precision.

Gangadhar et.al. Developed a tool for the query transformation [8]. The developed tool is flexible enough to correct the errors. N. Okazaki et.al. Present's a discriminative approach to generate candidate strings [9]. To generate candidate strings a string substitution rule is used and regularized logistic regression model is used to score these rules. Also the process is developed that works on the negative factors that degrades the performance across the precision boundary of the system.

A string transformation method used in this approach is tractable hence candidate strings can be enumerated. Experimental evaluation says that the performance of the system is efficient for the three tasks: lemmatization, noun identification and geographic variants. Tejada et.al. Advances the concept of active atlas [10] which is discussed in [11, 12]. In this paper main focus is given on string transformation weights. These weights are used to obtain the high degree of accuracy and precision in domain independent mania. As the system is domain independent it gains lots of attention in the area of query transformation.

Huang Weidong et.al. Done study on ontology for emergency retrieval of knowledge [13]. Whenever emergency retrieval of

knowledge is need to be done three problems may be arrives. 1. Organization of knowledge, 2. The expression to retrieve the documents, 3. Effective algorithms for retrieval. Hence the first problem can be overcome by two methods. 1. An ontology based database is generated and 2. Source of knowledge data is searched using internet. For another two issues there is need to increase the semantic precision.

Roberto et.al. Illustrates the ontology related information retrieval system used for arts and it known as WikiArt [14]. A WikiArt is a system that able to generates the result by considering three types of the sources i.e. ontology, wiki and a database Hence we can say WikiArt is a specialised system based on the Wikipedia that extract the results for the users that interested in Arts. WikiArt performance the semantic retrieval of documents hence gets higher precision when used with ontology.

K.Saruladh et.al Done survey on the Semantic Similarity Methods for Ontology base Information Retrieval [15]. The main purpose behind the survey is to check how exactly this similarity based methods are worked under ontology concepts. Basically there are number of similarity based methods are there but each of this method is falls under tree categories, information content, edge counting and node based counting.

Hao Wu, Guoliang et.al. Explains ginix: generalized inverted index for keyword search which speeds up the process of word searching [16].

3. PROPOSED SYSTEM

In this section, we describe our framework for String transformation approach using ontology with the below mentioned steps as shown in figure 1.

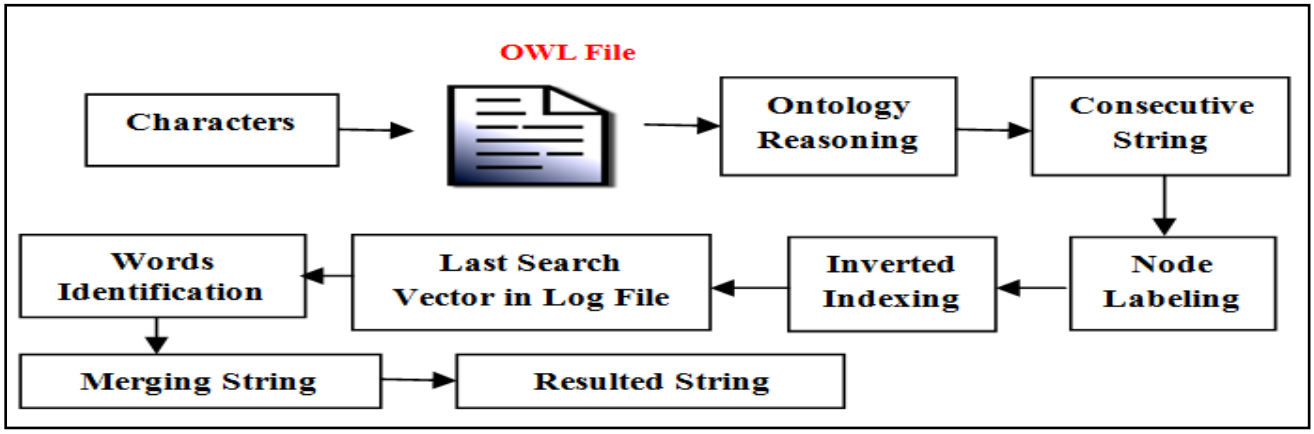

Fig 1: Overview of the proposed work

Step 1: In this step the characters are taken as input parameter for the proposed system and then they are passed to the ontology where reasoning is done to maintain the semantic of the word.

Step 2: Extensible Markup Language (XML) constitutes the syntactical foundation of the Semantic Web. So in this step OWL file which is been created by the protégé tool to access according to XML expression tags and create OWL property object.

Step 3: In this step we built classification Hierarchy, which was created using Reasoner in protégé. Protégé allows different OWL Reasoner to be plugged in; the Reasoner sent with Protégé is called Fact++. The ontology can be 'sent to the Reasoner' and automatically compute the classification hierarchy, and also check the logical consistency of the ontology. In Protégé 4 asserted hierarchy is the 'manually constructed' class hierarchy. Inferred hierarchy is the Class hierarchy that is automatically computed by the Reasoner. This is the most crucial part of our proposed system as our system retrieves all OWL properties according to the hierarchy of the OWL Reasoner.

System takes the OWL property object from the previous step and then identifies the parent and child classes. According to the hierarchy set by the protégé Reasoner our system classify the parent and child classes with their properties in the form of nodes. Finally these nodes are rearranged in a tree object.

Step 4: In this step, we use protégé tool to develop OWL ontology. Protégé is an open source, knowledge acquisition system, free ontology editor, and a framework for which various different projects propose plugins like Eclipse.

Protégé is developed at Stanford University, it can be extended for plug-in architecture, and this is a Java-based Application Programming Interface (API) which is developed for knowledge-based tools and applications. Protégé has more than 160,000 registered users.

The Protégé platform supports two important ways of personalization. The Protégé-Frames editor allows users to build and populate ontologies that are frame-based, as per Open Knowledge Base Connectivity protocol (OKBC). In this model, ontology can be created by a set of classes structured in a hierarchical type to represent a particular domain's main concepts. The Protégé-OWL editor allows users for creation of ontologies for the Semantic Web using W3C's Web Ontology Language (OWL). OWL ontology includes different descriptions for classes, properties and their instances. System takes the OWL property object from the previous step and then identifies the parent and child classes.

OWL Lite uses limited features of the OWL language. OWL Lite has many more limitations of the features than OWL DL and OWL Full languages. For example, in OWL Lite classes can only be used in terms of named super classes and only certain kinds of class limitation can be used. Equivalence in between classes and subclass relationship is allowed only in named classes not in arbitrary class expressions, and OWL Lite use only named classes. OWL Lite sub-language uses limited notion of the cardinality - the cardinalities accepted explicitly are 0 or 1. The following OWL Lite features related to RDF Schema are used in our approach.

- Class
- rdfs:subClassOf

- rdf:property
- rdfs:subPropertyOf
- rdfs:domain
- rdfs:range
- Individual

Step 5: An indexer system access the tree object of the Reasoner and then identifies the OWL class based on the successive characters of the query keyword.

Step 6: In this step last relevant query keywords will be searching from the log file using the generalized inverted index method to speed up the process as mentioned in the below algorithm.

Algorithm: Generalized inverted index

Input: Query vector $Q_{v,}$
User Entered Query as Q
Output: vector index

Step 0: start
Step 1: Get query Q.
Step 2: Divide query on space and put words in vector V.
Step 3: **for** i=0 to length (V).
Step 4: **for** j=0 to size of Q_v
Step 5: Get a document D_j
Step 6: **if** V_I € D_j **then**
Step 7: add D_o into set G
Step 8: **End inner for**
Step 9: **End outer for**
Step 10: **For** i=0 to length of G
Step 11: **if** Gi and Gi+1 **then**
Step 12: Add interval List in set L
Step 13: **End FOR**
Step 14: **for** i=0 to length of L
Step 15: add beginning and END index of L_i into set I_n(inverted index)
Step 16: **end for**
Step 17: Twin Heap (I_n)
Step 18: Stop

Step 7: In this step the key words which are found from the log files are compared with ontology words and finally the resulted words are merged to show the most probable expecting words from the user.

Mathematical Model

The whole proposed system is expressed mathematically in the below model.

1. Let S={ } be as system for String transformation using ontology

2. Identify Input as Q={ Q_1, Q_2, Q_3…..Q_n}

   Where Q_n= Number of input character for the query word

3. Identify A as Output i.e. probable answer word

   S= {Q_n, A}

4. Identify Process P

   S= {Qn, A, P}
   P= {Qn, A, O_r, $L_{l,}$ G_i}

   Where
   Q_n = Query keyword characters
   O_r =Ontology reasoning
   L_l=Log linear model
   $G_{i=}$ Generalized inverted index
   A= Answer word

5. S = { Q_n, A, O_r, $L_{l,}$ G_i }

6. Ontology Reasoning

   Set Or:
   Or_1 = Read the OWL file
   Or_2= Create hierarchy tree
   Or_3= Find right child
   OR_4=Load all children in a vector

7. Log linear model

   Set Ll:
   Ll_1=Get the log path
   Ll_2= Read the log file
   Ll_3= Get line content in a vector
   Ll_4=Get the occurrence line index
   LL_5= Create a keyword vector index

8. G_i Generalized inverted index

   Set $G_{i:}$
   $G_{i1=}$ Get the vector index
   $G_{i2=}$ Identify the index status in inverted form
   $G_{i3=}$ Perform twin heap
   $G_{i4=}$ Sorting index
   $G_{i5=}$Identifying the word

The union of all subset of S Gives the final proposed system.

4. RESULTS AND DISCUSSIONS

To show the effectiveness of proposed system some experiments are conducted on java based windows machine using netbeans as IDE. To measure the performance of the system we set the bench mark on different number of keywords for many input queries for designed sample searching system. As the input characters are entering the system automatically triggers the working process by reasoning the ontology OWL file and search the proper keyword based on the linear log model which is catalyzed by generalized inverted index system. The experimental results are plotted the below figure 5.To measure the performance of the system we uses WAPT 7.0 tool in web paradigm.

Table 2: Values for graphical result

No. of users	10	20	30	40	50
Performance Time in Sec	4	7	9	11	12

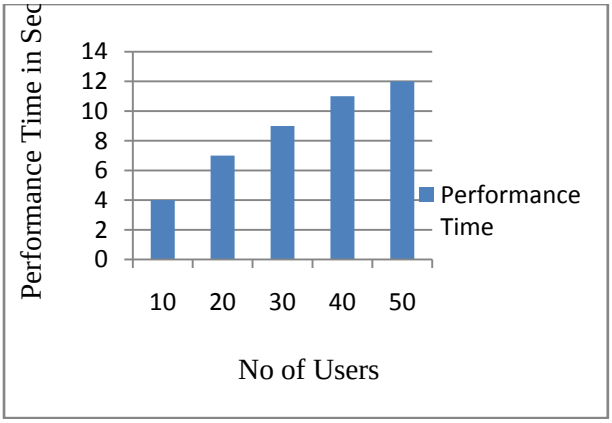

Fig 2: Performance measurement for different no of users.

The above plot expresses result of String transformation performance time which is not directly proportional to number of users. So this indicates that the system over performs the string transformation system in matter of time.

5. CONCLUSION AND FUTURE SCOPE

Proposed method successfully applies the String transformation method on searching systems using the ontology and generalized inverted index method. Here in our model proper ontology reasoning will be done to identify the best hierarchy of the keyword in the tree of the words without losing their semantics. Again for searching similar word from the log file linear log model is applied where system linearly searching the keywords. To accelerate this process system successfully weave the model of generalized inverted index system to fasten the searching process. So that system will yield the result in best possible time.

The proposed system can be enhance to consider more bag words in either from cloud or distributed systems to provide the best possible probability for the answers.

6. ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude and deep regards to Prof. S.R.Todmal for his guidance and continuous encouragement throughout.

7. REFERENCES

[1] 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.

[2] Huizhong Duan, "*Online Spelling Correction for Query Completion*",ACM,WWW 2011,March28-April 1,2011,Hydrabad, India..

[3] Rishabh Singh, "*Learning Semantic String Transformations from Examples* ", Microsoft Research, Redmond,2011.

[4] Miguel R. Luaces, Jose R. Param´a, Oscar Pedreira, and Diego Seco,"*An Ontology-based Index to Retrieve Documents with Geographic Information*",ISO/IEC: Geographic Information,2002.

[5] Sajendra Kumar, "*Ontology based Semantic Indexing Approach for Information Retrieval System*", International Journal of Computer Applications, Volume 49– No.12, July 2012.

[6] Thanh Tran, Philipp Cimiano, Sebastian Rudolph and Rudi Studer "*Ontology-based Interpretation of Keywords for Semantic Search*",X-media(IST-2006- 026978) and the NeOn (IST-2006-027595) projects.

[7] Jeyalakshmi.S, Rathika.T "*Improving Efficiency And Accuracy In String Transformation On Large Data Sets*",International Journal of Computer Science & Mobile Application, Vol.2, Issue 3,March-2014,pg.55-65.

[8] Gangadhar "*A Tool for String Transformation Enabling Flexibility in Real World Applications*", International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 11, November 2014.

[9] 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 '2008,pg.447-456.

[10] S.Tejada, C. A. Knoblock, and S. Minton, "*Learning domain independent string transformation weights for high accuracy object identification*" 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] S. Tejada, C. A. Knoblock, and S. Minton.,"*Learning object identification rules for information integration*", Special Issue on Data Extraction, Cleaning, and Reconstrction, Information Systems Vol. 26, No. 8, pp. 607-633, 2001.

[12] 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.

[13] Huang Weidong, Zhu Hengmin, Wang Chuanlei, Shen Yu,"*Research on Emergency Knowledge Retrieval Mechanism based on Domain Ontology*",IEEE,Intelligent Systems,May 2009,pg.212-216.

[14] Roberto Pirrone, Vincenzo Cannella, Orazio Gambino, Arianna Pipitone "*WikiArt: An Ontology-based Information Retrieval System For Arts*",IEEE,Intelligent System Design & Application,ISDA,2009.

[15] K.Saruladh,"*A Survey of Semantic Similarity Methods for Ontology based Information Retrieval*",IEEE,Second International Conference on Machine Learning and Computing,2010, pg.297-301.

[16] Hao Wu, Guoliang Li, and Lizhu Zhou,"*Ginix: Generalized Inverted Index for Keyword Search*", Ieee Transactions On Knowledge And Data Mining Vol:8 No:1 Year 2013.