TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

A Critical Review of the Unguided Loose Search (ULS) Process for Natural Language Based Extraction Technique on Relational Databases

[1]Enikuomehin A.O., [2]Sadiku A.S., [3]Egbudin M.D
[1]Department of Computer Science, Lagos State University, Lagos Nigeria
[2]Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[3]Infinity IT, Lagos Nigeria , Nigeria
toyinenikuomehin@gmail.com, assadiku@unilorin.edu.ng, dechyzeay@yahoo.com

ABSTRACT

Formulation of query statements by searchers for submission into relational databases and information retrieval systems have been a serious challenge that often lead to irrelevant search results. This is compounded by the level of uncertainty about the user's information need and in some cases, unfamiliarity with retrieval system. Evidently, the World Wide Web presents a more established challenge in this area, considering the fact that searchers has little or no training on search techniques on the web. This paper recognizes fuzzy logic system and fuzziness as a tool required to close the gap between automated systems and human thinking. We realize this stiffness in query presentation as against the flexibility in human thinking and then consider the fuzzy concept as a tool that can be incorporated into a new system to overcome the syntactic problem presented in most relational operations. Thus, the paper proposes a novel approach of natural language query based problems. We propose the use of an Unguided Loose Search (ULS) which involves the use of local appropriator on a fuzzified Natural Language Interface. Our approach incorporates fuzziness in the interface, using the local appropriator, of the database systems rather than within the data itself. It allows freedom to users since they will not have to learn any specialized syntax such as that of SQL. The result shows that the new querying model called the EFUSQL model is applicable to real life users and can be incorporated into existing databases and query interfaces. The results show that naïve users prefer the new system due to its flexibility and response time.

1 INTRODUCTION

Human are vague in nature. Incomplete, Imprecise and uncertainty that result from users intention (vagueness) for querying databases has largely not be handled with the existing querying system since queries are not discriminated. There is an extensive research examining

DOI: 10.14738/tmlai.24.308
Publication Date: 3rd Aug 2014
URL: http://dx.doi.org/10.14738/tmlai.24.308

how imprecise and uncertain data can be presented in, and queried from databases given that it is pervasive in most real-world applications. Data extraction, document retrieval, query execution and answering etc have been extensively studied. It has mostly involved the transformation of an information need to a syntactical form in terms of query, in this the user need to have a prior knowledge of the database domain. If we consider as example, the search for the age of a student in a particular university, we will need to formulate a query that will have a target table on a particular database. Problem of execution exist when the user does not know the database table name. It means that the degree to which the query represents a user's information need is a function of user's ability to precisely formulate the information need in a suitable syntactical form admissible within the sql syntax formulation domain. Similarly, many research efforts have been carried out to extend the database models and query languages in other to incorporate fuzzy representation and query capabilities for fuzzy data, the argument still remains that, it is challenging to present a system that is useful and practicable to users since research has been on querying fuzzy databases with some level of fuzziness whereas most real databases are not fuzzy and database administrators still prefer the use of crisp database as evident in most organizations. Furthermore, while some users may want to continue using the traditional facilities available in most of the key commercial database systems and not interested in learning new query languages, other groups of users may want to use fuzzy linguistic terms for querying a non-fuzzy database. This is the case with sql and its extentions. First, users need to know about the existence of the data and secondly, they need to be equipped with technicalities to retrieve such data. This work concentrates on the latter task. Today, one of the biggest challenges in web technologies is data retrieval in multidimensional databases. A multidimensional database is a database that hold data as text in document, Images, video etc, it is a relational database with external links with other databases. To cope with this information growth, existing search methods will need to be enhanced to appreciate an acceptable level of relevance. Query results do not satisfy the users to a large extent thus users are forced to make a decision or a choice based on the displayed output. The same problem occurs also within an organization when a staff is searching for a data from a single database. This can best be explained by the incident of the December 25th 2009 bombing attempt of a Detriot, USA based flight by a Nigerian. In that incident, which can attributed to Query misrepresentation, Adam Brookes released a bulletin on the said bombing of the flight 253 "Once again, it is the failures of the US intelligence agencies that, we are told, are to blame. The report found out that the US government did have 'sufficient information' to disrupt the Christmas day attack. But that information was scattered around databases. It was never pulled together to present a coherent picture of the threat. A 'series of human errors' occurred, apparently someone misspelled Umar Farouk Abdulmutallab's name as they entered it in a database and that is why no-one realize he had a US visa. The above is a clear indication of the problems this paper attempts to resolve. These could exist in a singular database or also a multidatabase system evenly spread and distributed on the internet. Therefore , the major

context here is not the web usage but the perfect method and tools for extraction of the open and hidden data. In most cases, these data are available, however the tools such as the query language, need to be sufficiently enhanced to be able to provide a human usable result. This clearly exceed the capability of the existing database querying tool ' THE SQL and its extentions' . Web interfaces to databases are relatively simple and restricted. Even a skilled user could not define complicated queries due to their limitations. therefore, databases that need to be accessed via the Web should offer more "intelligence". In this paper, we are shifting the intelligence from the database to the query since most relational database users still use the conventional databases and not much of fuzzy databases or any other form of databases that has so much be covered in literature.

2 FUZZINESS IN DATABASES

2.1 Theory of Fuzzification

The original interpretation of Fuzzy sets arises from a generalization of the classic concept of a subset extended to embrace the description of "vague" and "imprecise" notions. This generalization is made in he following way:

1. The membership of an element to a set become a "Fuzzy" or "vague" concept. In the case of some element , the issue of whether they belong to a set may not be clear.
2. The membership of an element may be measured by a degree, commonly known as the "memebership degree" of that element to the set, and it takes a value in the interval [0,1] by agreement.

Using classic logic, it is possible to deal only with information that is totally true or totally false; it is not possible to handle information inherent to a problem that is imprecise or incomplete, but this type of information contains data, which would allow a better soution to the problem. In classic logic the memebership of an element to a set is represented by 0 if it does not belong and 1 if it does, having he set {0,1}. On the other hand, in Fuzzy logic this set extends to the interval [0,1]. Therefore, it could be said that Fuzzy logic is an extension of the classic systems. Fuzzy logic is the logic behind approximate reasoning instead of exact reasoning. Its importance lies in the fact that many types of human reasoning, particularly the reasoning based on common sense are by nature approximate.

Note the great potential that the use of membership degrees represents by allowing something qualitative (Fuzzy) to be expressed quatitatively by means of the membership degree. A Fuzzy set can be defined more formally as follows:

Defination 1:

A **Fuzzy set** A over a universe of discourse X (a finite or infinite interval within which the Fuzzy set can take a value) is a set of pairs

$$A=\{\mu_A(x) \,/\, x : x\in X, \mu_A(x)\in[0,1]\in\Re$$

OR (1)

$$\mu_A(x)=\begin{cases}1, iff\ x\in A\\ 0, iff\ x\in A\end{cases}$$

Where $\mu_A(x)$ is called the memebership degree of the element x to Fuzzy set A. This degree ranges between the extremes **0** and **1** of the dominion of the real numbers:

- $\mu_A(x)=0$ indicate that x in no way belong to the Fuzzy set A
- $\mu_A(x)=1$ indicates that xd completely belongs to the Fuzzy set A

Sometimes, instead of giving an exhausive list of all the pairs that make up the set (discreet values), a definition is given for the function $\mu_A(x)$, referring to it as characteristic function or **membership function.**

2.2 Fuzzy sets Theory and Query processing in Databases:

In written sources, we can find a large number of papers dealing with this theory, which was first introduced by Lofit A. Zadeh in 1965 (1). A more modern synthesis of fuzzy sets and their applications can be found in (2), (3), (4) , (5), (6) .

The original interpretation of fuzzy sets arises from a generalization of the classic concepts of a subset extended to embrace the description of “vague” and “imprecise” notions. This generalization is made considering that the membership of an element to a set becomes a “fuzzy” or “vague” concept. In the case of some elements, it may not be clear if they belong to a set or not. Then, their “membership degree” of the element to the set, and it takes a value in the interval [0,1] by agreement. Using classic logic, it is only possible to deal with information that is totally true or totally false; it is not possible to handle information inherent to a problem that is imprecise or incomplete, but this type of information contains data that would allow a better solution to the problem.

Querying is the process of retrieving information or data from the database. The traditional query in a relational database has been shown to be incapable to satisfy the needs for dealing linguistic values. The structured query language Sql has been around for while from RDBs, the Sql is a declarative language that allows the user to specify “what” information from the database is needed without having to specify how it is to be retrieved: IE is constructed such that each DBMS translate individual query into an efficient execution plan. Recall that these were issue of bi-valued interest if an item belongs into a set or not since the time of Aristotle’s there. The answer has been usually formed as a simple truth function assuming only values YES or NO for an answer. A thought shift has been made in 30’s of the last century particularly by Lakeview 2’s thesis. The contemporary SQL norm supports only classical bi-valued logic. Unfortunately, the use of fuzzy sets and fuzzy logic operations is not defined and there are

many of mutually different commercial and General Public License SQL server distribution in essential SQL norm implementation.

3 APPROACH

3.1 Retrieval based on Similarity

The approach used is an extension of the concept of retrieval based on similarity as a function of relevance. This is used in the formulation of an appropriate model as a benchmark for the work. Relevance is measured by concept of similarity. There are two type of relevance in similarity:

3.1.1 Fuzzy similarity

Definition 2 .

Let L be a Fuzzy algebriac function and let A be a non-void set. A fuzzy similarity S on A is such a binary fuzzy relation that, for each x, y, and z in A,

i. $S(x,x) = 1$ (everything is similar to itself),

ii. $S(x,y) = S(y,x)$ (fuzzy similarity is symmetric),

iii. $S(x,y) \circ S(y,z) \leq S(x,z)$ (fuzzy similarity is weakly transitive)

3.1.2 Data Similarity

For every value 't' in the domain of attribute 'A', D(t) can be defined as $\log(n/F(t))$,

where 'n' = number of tuples in the database

F(t) = frequency of tuples in database where 'A' = 't'

The similarity between a tuple 'T' and a query 'Q' is defined as: i.e., similarity between a tuple T and a query Q is simply the sum of corresponding similarity coefficients over all attributes in T.

3.2 Computing top-k Answers

Assume a query Q with m elementary conditions on the attributes A_i, i in {1,…,m}. The multidimensional database D consists of a single relation R with a finite set of N tuples described on the attributes $A_1,\ldots,A_m$. Each tuple t is associated with a vector $(x_1,\ldots,x_m)$ of m scores, one for each attribute of the elementary query condition. Scores are computed from attribute values of each tuple with respect to their similarity to the query condition. For the top-k problem, the database could alternatively be seen as a set of m sorted lists L_i of N pairs (t, x_i), t in R. Hence, for each elementary condition of the query Q, there is a sorted list L_i in which all N database tuples are ranked in descendant order. Entries in the lists could be accessed randomly from the tuple identifier or sequentially from the sorted score. The main issue for

top-k query processing is then to obtain the k tuples with the highest overall scores computed according to a given aggregation function $agg(x_1,\ldots,x_m)$ of the attribute-oriented scores x_i. The aggregation function agg() used to combine elementary conditions has to be monotone; that is, agg() must satisfy the following property:

$agg(x_1, \ldots, x_m) <= agg(x_1',\ldots, x_m')$ if $x_i <= x_i'$ for every i. (2)

Among the various monotone aggregation functions are t-norms and t-conorms respectively associated with conjunctive and disjunctive queries, and weighted means as well. Min and Max are the most common functions respectively for conjunctive and disjunctive queries. The naive algorithm consists in looking at every entry (t, x_i) in each of the sorted lists L_i, computing the overall grade of every object t, and returning the top answers. Obviously, this approach suffers from a high access cost to the lists since all the N overall grades are computed.

Let us address set of all N tuples in a relational database table as the tuple table, sorted by the decreasing order of score. We store information about the x-tuples in an x-table. By using a hash map, given the id of a tuple t, the score and confidence values for all its alternatives can be retrieved efficiently from the x-table in O(1) time(Worst Case).

To process a top-k query, we retrieve tuples in decreasing score order and stop as soon as we are certain that none of the unseen tuples may possibly affect the query result. The dept of the search denoted by n can be defined as the minimum number of tuples retrieved so as to generate the correctness of the result. The k top approach can be used to fomalize a model presented below:

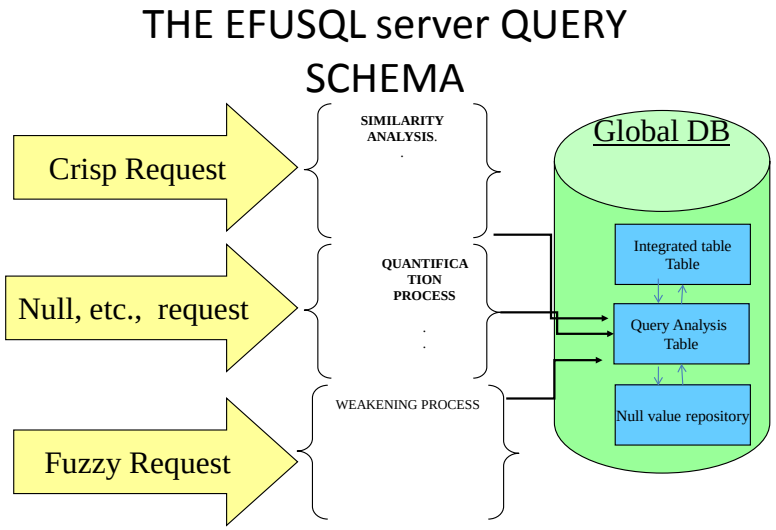

Figure 1: A similarity system on a multidimensional DB

4 EXPERIMENT

An application interface developed in php is placed over the sq1 server for the purpose of translation. We did not use the Fsq1 server of medina as it has not been fulfilled practically and more so 77% of RDBMS users still use the conventional sq1 server and no Fsq1 server. It acts as a middle ware that translates the National Language to crisp sq1. Consider the query:

Brilliants students that are young:

Pre query analysis:

"Student" is the key word; therefore it is awarded a relevance 'Mark R'.

The system analyses the query by gathering information from different data sources, combine them to a standard format then store on a location. The associated sample table is shown in table 1.

Table 1: Database table to model performance

NAME	CGPA	LEVEL	STATE OF ORIGIN
Toyin	2.70	300	Ondo
Usman	3.40	400	Kano
Smith	2.50	300	Lagos
Kunle	4.00	200	Kwara
Shade	2.67	100	Oyo
Nnena	3.35	200	Imo
Obiora	2.21	200	Anambra

In the process of establishing a truly fuzzy querying system, other information than the crisp inserted data may be required: As example, for state of origin, one way need local government area, or village name or family house. This other information are called Meta information since they provide more information about the data. The introduction of metadata in commercial search engine is a novel idea whose popularity has not been fully harnessed. Its importance is in its ability to provide non discrete information about data. The meta table contains all the information required for fuzzification in stage (3). The meta-table is given as follows:

Linguistic –Term: used to store the name of the fuzzy set.

Table- name: used to refer name of the table in which attribute associated with the fuzzy set is available.

Column-name: used to refer to attribute associated with fuzzy set.

Alpha (α): lower range of the SUPPORT [Ross, Fuzzy logic in Engineering]

Beta (β): lower range of the CORE [Ross, Fuzzy logic in Engineering]

Delta (δ): upper range of the CORE [Ross, Fuzzy logic in Engineering]

Gamma(γ): upper range of the SUPPORT [Ross, Fuzzy logic in Engineering]

The core of a membership function for some fuzzy set A is defined as that region of the universe that is categorized by a complete and full membership in the set A. that is the CORE comprises those elements x of the universe such as that $\phi_A(x) = 1$.

SUPPORT the region that is characterised by non zero membership in the set A i.e. $\phi_A(x) > 0$.

The core of a membership function for some fuzzy set A is defined as that region of the universe that is categorized by a complete and full membership in the set A. This approach is different from the earlier approaches as in (7), (8) where membership functions wer not used, however thresholds were included. At the end of it, after membership fixes are generated, the values are now defuzzified and ranked in ascending order.

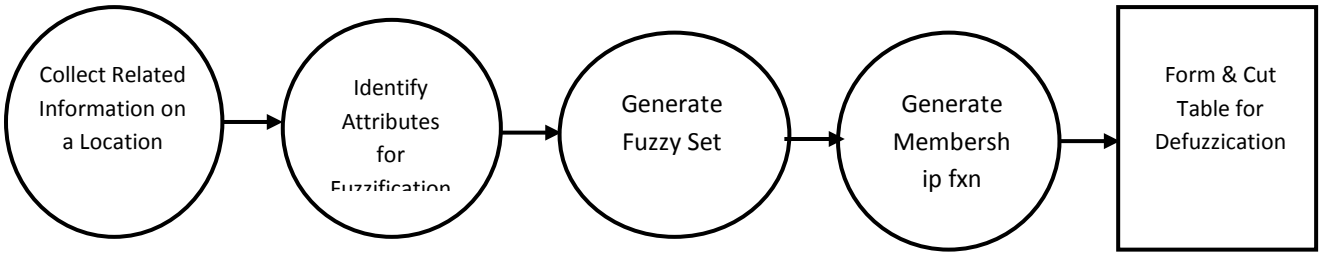

Figure 2: The data data fuzzification flow

5 COMPUTATION OF MEMBERSHIP FUNCTIONS

To allocate membership value, the fuzzy sets generated is divided into three groups;

Group A:

The set of terms and values at the lower boundary region.

Group B:

Th The set of terms and values at the is forms the core region

Group C:

The set of terms and values at the upper boundary region.

The membership value for the core region is always to 1. The lower boundary and upper boundary are calculated as

$$(x- \alpha)/(\beta - \alpha) \; and \; (\delta - x)/(\delta - \gamma)$$

where x is the value of the attribute that can be brought from the concerned table.

To incorporate weight, we consider a fuzzy relation R such as shown in table 2 below and called the grading factor used in generating the values in table 1:

Table 2: Grading Factors

		Excellent	Very Good	Good	Fair	Poor
	Correct answer	0.3	0.4	0.3	0.1	0
R =	Language	0.0	0.2	0.5	0.3	0
	Presentation	0.1	0.6	0.3	0	0

Now, the professor want to assign a grade to each paper, we formalize this approach, thus Let X be a universe of factors and Y be a universe of evaluations, so

$$X = \{ x_1, x_2, \ldots\ldots\ldots x_n\} \text{ and } Y = \{y_1, y_2, \ldots\ldots\ldots\ldots\ldots y_n\} \quad (4)$$

where i = 1, 2, … n and i= 1, 2, … m.

Suppose we introduce a specific paper into the evaluation process in which the professor has given a set of "scores" (w_i) for each of the n grading factors, we ensure, for convertion, that the sum of the scores is unity. This means that each of the scores is actually a membership value for each of the factors xi, and they can be arranged in a fuzzy vector $\underline{\omega}$, to have,

$$\underline{\omega} = \{w_1, w_2, \ldots\ldots\ldots, w_n\} \quad (5)$$

where

$$\sum_i w_i = 1$$

The process of determining grade for a specific paper is equivalent to the process of determining a membership value for the paper in each of the evaluation categories, y_i. This process is implemented through the composition operation

$$\underline{e} = \underline{\omega} \text{ o } \underline{R}$$

where e is a fuzzy vector containing the membership values for the paper in each of the y_i evaluation categories.

After the generation of the membership function and the weight, we compute the membership values corresponding to the attribute. To do this, we propose the system checks for hedges (fundamental atomic term are often modified with adjectives (nouns) or adverbs (verb) like very low, slight, more or less, fairly, almost etc, that is, the singular meaning of an atomic term is modified or hedged) from its original interpretation. Using fuzzy set as the calculus of interpretation those linguistic hedge have the effect to modify the mf for a basic atomic term (9). When the hedges is calculated and manipulated, the next stage removes the fuzziness contained in the query by the use of appropriate deffuzzication technique. However, in our model, we introduce a Hard Aggregation Concepts (HAC) before the deffuzzication processes take place. In the Hard Aggregation Concepts (HAC), other components of the query are

segmented in parts. These parts are the Modifier Part (MP) and the Concepts Part (CP). We propose

$$MP + CP = HAC \tag{6}$$

In practice, there might be more than one HAC in a query, then we say,

let there exist HAC such that any term in query is an element of the HAC, representing the i=th HAC. To explain our idea of the HAC, let use the following example;

Find the name, level, age of very young and quite tall students where grade ≥ 4.0.

The HAC can be implemented as follows;

[name, age, grade] are return attributes

[very, quite] are Modifier Parts

[Young, tall] are Concept Parts;

Student is a table on the database

Grade ≥ 4.0 is a Crisp Condition.

After this "Classification", the above computation of membership function is carried out on the HAC components and the values are now integrated into the query range. Then deffuzzification is then implemented by finding the α - Cut and by calculating the maximum and minimum range. Once of minimum and maximum range is calculated with the fuzzy terms with linguistic hedges are remodel and the result displayed. This means that the central meaning of the query has been taken into account in the query processing. This work introduces a new process of implementing fuzzy queries with the use of membership value manipulation and HAC. It enables us to write natural language fuzzy queries at frontends and get discriminated results, this also helps in handling missing data since if a row satisfies at least one fuzzy criteria or crisp criteria, it will be accommodated in the range and then will be included in the result set even if the data has some missing attributes.

An important aspects of this research is that this query system can process multiple fuzzy queries in plain in human language as well as crisp query at the same time with optimum intelligent result. Steplan et al (01) presents a Skyline operator for air flight selection, which select best rows or all non-dominated based on a crisp multi-criteria comparism. A row dominates the other if it is as good or better than the other in all multiple criteria and better in at least one criterion.

6 CONCLUSION

In the implementation of the suggested technique, the followings significant observations were made; the query language used is highly flexible i.e. user need not bother about the syntax of the language, queries may contain fuzzy, uncertain and imprecise terms. Database schemas

need not be modified for storing the membership of individual record, Fuzzy extensions are being created automatically so overhead of neither user nor DBA is being increased, Updating of fuzzy extensions after each and every update in master database is also being done automatically with the help of triggers, In the implementation of this technique only logical view of database is being affect(Table referencing), Crisp queries are also being handling along with the fuzzy ones. Intelligence has now been incorporated with searching to get result much more human. The paper, following these conclusions recommends that non crisp processes should be included in the highly patronized commercial database.

REFERENCES

[1]. Zadeh, L.A. "Fuzzy sets". Information and control, 8, 1965, pp. 338-353

[2]. Buckley, J.J., & Eslami, E. "An Introduction to fuzzy logic and fuzzy sets" (advances in soft computing). Physica-Verlang Heidelberg, 2002

[3]. Kruse, R., Gebhardt, J.,& Klawonn, F. "Foundations of fuzzy systems". John Wiley & Sons, 1994

[4]. Mohammad, J., Vadiee, N., & Ross, T.J.(Eds.). "Fuzzy logic and Control: Software and Hardware applications". Eaglewood Cliffs, NJ: Prentice Hall PTR, 1993,

[5]. Nguyen, H.T., & Walker, E.A. "A first course in fuzzy logic (3rd ed.)". Chapman & Hall/CRC, 2005, Piegat, A. "Fuzzy modeling and control". Physica-Verlag (Studies in Fuzziness and Soft Computing), 2001

[6]. Pedrycz, W., & Gomide, F. "An introduction to fuzzy sets: Analysis and design" (A Bradford Book). The MIT Press, 1998].

[7]. Galindo, J., Medina, M., Pons, O., & Cubero, J. (1998). A Server for Fuzzy SQL Queries. In T. Andreasen, H. Christiansen, & H. Larsen (Eds.), Lecture Notes in Artificial Intelligence (Vol. 1495, pp. 164-174). Springer.

[8]. Bosc, P., & Pivert, O. (1994). Fuzzy Queries and Relational Databases. Proceedings of the 1994 ACM Symposium on Applied Computing, (pp. 170-174). Phoenix,AZ.

[9]. Galindo, J., Urrutia, A., & Piattini, M. (2006). Fuzzy Databases: Modeling Design and Implementation. Hershey:PA: IDEA Group.