Dynamic Fuzzy String-Matching Model for Information Retrieval Based on Incongruous User Queries

Olufade, F. W. Onifade, *Member IAENG,* Odile Thiéry, Adenike, O. Osofisan, Gérald Duffing

Abstract - Information representation and reasoning strategies are important aspects of information retrieval that facilitates adaptability. Unfortunately, most research focus has been on the representation, leaving the users' information need at the mercy of the system. Reasoning strategies assists with some level of intelligence in adapting user queries to the content of the database, thereby reducing the possibilities of "no match" in the face of available information. Query incongruity can result from uncertainty, misinformation, misrepresentation, lack of adequate knowledge in a domain, and phonological transposition amongst other. A particular focus is meeting the information need of the dyslexics. We proposed a Fuzzy String-Matching model to tailor and assist in information retrieval resulting from incoherent user queries. Our model fared well in assisting user's access to desired information consequent on the fuzzy reasoning model strategy.

***Index Terms*—Fuzzy String-Matching, Fuzzontology, Information Representation, Information Retrieval, Reasoning Strategy,**

I. Introduction

The exponential growth in the available information witnessed in the last decade has resulted in the proliferation of information retrieval objects, which consist of algorithms, methods, technologies and tools. These objects are saddled with the responsibility of ensuring user access to prompt and adequate information, however, the story is not always as expected. Information retrieval (IR) is the scientific discipline concerned with the analysis, design and implementation of computerized systems aimed at addressing the representation, organization of, and access to vast amount of heterogeneous information already encoded in digital format [21] .

The expectation from any Information Retrieval System (IRS) is to make available such information considered pertinent to a user's query (formally expressed in the system's query language). Unfortunately, these goals are not deterministic sequel to the presence of uncertainty and vagueness embedded in many parts of information retrieval process [7]. Canfora & Cerulo [5] opined that a key feature of an IRS is the retrieving the document satisfying the information need of a user from amongst huge collection of documents. These systems, in web context are referred to as search engines. In a bid to facilitate ease of search, user information request are represented by keywords or phrases that are indexed. These representations are known as queries, and the indexing can assume diverse terms depending on the tools, however, ranked IR methods are popular i.e. documents are ranked based on measure of relevance as compared to user's request.

The expectation is thus on the IRS to appropriately deal with the concept of uncertainty and vagueness which has been majorly ignored in commercial IRS. Another user expectation is the expected flexibility in IR process. Flexibility in this regards implies the capability of the system to manage imperfect (vague and/or uncertain) information, and also to adapt its behaviour to the user context. Centrally, the main goal in IR is the quest to the set of *relevant* documents, amidst large collection in a bid to satisfy the information need expressed in form a query by a user [8]. We note that these large documents can be in form of texts, images, video, audio, mediums, or sometime in a multiple format of any combination of the above mentioned but the focus is on text documents.

The rest of the paper is organized as follows: we take a look at information retrieval and search engines' operation in section II. Section III comprises of information representation and reasoning strategies. We introduced our model in section IV with real life examples and section V concludes the work.
Procedure for Paper Submission

II. Information Retrieval & Search Tools

Search engines are resources to assist users in information retrieval. Information retrieval is inherently predicated on users searching for information from their "information need" that result from the interpretation of the decisional problem. Apart from the quality of data and information in the data warehouse, the volume, timeliness of the information to the decision maker is equally important. While the information need might be right, inherent errors resulting from dirty data are detrimental to the overall goal of information retrieval.

Since most search operations are performed on the internet or corporate organizations expensively constructed and maintained data warehouses, [4] submitted that the main

Manuscript received February 25, 2010. This work was supported in part by French Ministry of Foreign Affairs (Nigeria).

Olufade, F. W. Onifade is with Equipe SITE - LORIA, Campus Scientifique, Nancy, France & Department of Computer Science, University of Ibadan, Nigeria. Phone: +33631765305, +2348074010558; e-mail: onifadeo@loria.fr).

Odile Thiéry. Equipe SITE - LORIA, Campus Scientifique, Nancy, France & Universite Nancy 2, Nancy, France. (e-mail: thiéry@loria.fr).

Adenike, O. Osofisan is with Computer Science Department, University of Ibadan, Oyo State, Nigeria, (e-mail: nikeosofisan@gmail.com).

Gérald Duffing is with Equipe SITE - LORIA, Campus Scientifique, Nancy, France & ICN Business School, 13, rue Michel Ney, 5400, Nancy, France (e-mail: duffing@loria.fr)

difference with the classic model for IR and the one augmented for the web lies in the replacement of "Matching Rules" in the former with "Search Engines" in the latter. This is shown in figure 1 below.

[4] linked information need with some set of tasks. Information need is usually verbalized (silently, mentally, and not necessarily loud) and this is translated into query submitted to the search engines. This information need determine the nature of the queries submitted for selection from a collection of documents (corpus) based on the matching rules. This background allows recognizing different sets of users. A user can either be experienced, versed, or inexperienced. The manner via which each of these users constructs their query goes a long way to determine the probable result from the databases.

In the area of cognitive model for web search [13] and [11] explore user's mental model for search engines with other related results presented in [6]. In all these models, there is an agreement that web searches are sequel to user's information need.

There are established models of IR which are the Boolean, Vector space, Probabilistic, and Fuzzy models. Various implementations of these models are in existence, but more and more surfaces as their limitations become apparent for retrieval purposes. Another popular approach to IR is based on the method of analysis of natural language [20]. It was however found out that this method is limited in the level of deepness of the analysis of the language, and their consequent range of applicability i.e. *satisfying interpretation of the documents' meaning needs a too large number of decision rules even in narrow application domains* [7] . Inexact string matching has also been adopted in many realms [24] but the flexibility of this method is not appropriate in vague, uncertain and dynamic environment.

The main components of IRSs are: collection of documents, a query language allowing the expression of selection criteria synthesizing the user's needs, and the matching mechanism which estimates the relevance of the documents to the query [16]. Attempt to estimate the relevance of each document with respect to a specific user need is based on a formal model which provides a formal representation of both documents and the user queries. Using the trio of documents collection, query language and the matching mechanism in an IRS, the input represents the user's query while the corresponding output reflects the relevance estimation of the user information need (query) and the information collection.

Apart from the query language, of importance is the representation of the document's information content, this takes the form of keyword extraction and weighing. It is however worth noting that documents representation are done without taking into cognizance the subjective view of the users on the documents [2]. Attempt at making sure that a user retrieve documents "*relevant*" to her query necessitate a formal representation of the documents contents known as "*Indexing*".

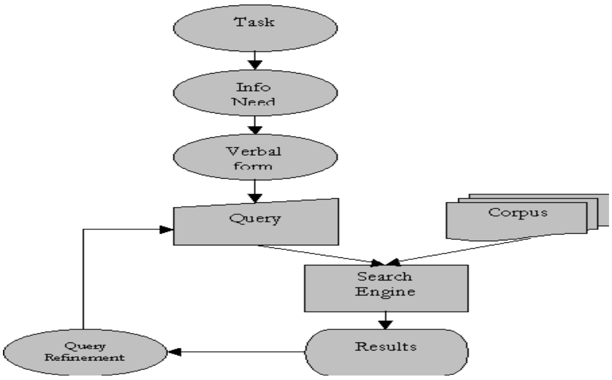


Fig. 1: Web-Augmented Classic Information Retrieval Model [4]

With this method, analysis of a document is followed by a surrogate describing the document in the index. With this in mind, a query to an IRS provides either an exact answer or a ranking of document with highest possible relevance [9]. The result thus is a function of the formal model adopted in designing the system.

III. Information Representation and Reasoning Strategists

Modeling the process of information retrieval is generally complex consequent upon the fact that it is multifaceted and inherently endowed with vague concepts difficult to formalize. Human component has been of focus lately considering the issue of relevance, information need and other subjective factors. The importance of reasoning strategy employ alongside the representation of information has necessitated the improvement made on figure 2. Reasoning strategy facilitates the representation of similarity problem in computing the relevance of a document with regards to submitted queries [5].

Our first consideration is the representation of an information retrieval model by [12]. Information retrieval model was characterized by a set of quadruple given as {D, Q, F, R(q, d)} and these factors were defined as follows:

- D is a set of *logical views* for the documents in the collection, it is a *representation component;*
- Q is a set of *logical views* for the user information needs, it is a *representation component;*
- F is a *framework for modeling document representation*, *queries* and *their relationships*, it is a *reasoning component*;
- R(q, d) is a *ranking function* which *associates a real number with a query q* Q and *a document d* D. It is a *reasoning component*.

The disparities in consideration of both components of information retrieval models have been the basis for the lopsided query results for users' information needs. The available tools for retrieval is based majorly on representation and less on reasoning, thus result for the query will be based on either being relevant or irrelevant and not necessarily accommodating some vagueness and uncertainty in the retrieval operation (fig. 2).

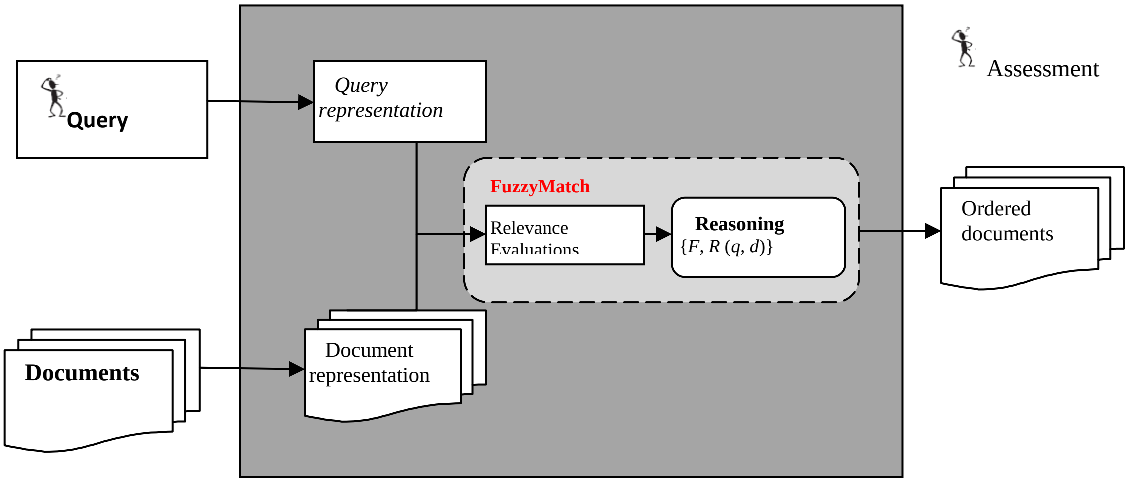


Fig. 2: Proposed model for Information retrieval

IV. Model Descriptions

The need for accommodating vagueness and uncertainty in database representation has been the basis for the introduction of fuzzy systems into the field. Popular soft information retrieval models are based on fuzzy set theories and the connectionist (neural networks) theory. Considering the Boolean query where achieves are partitioned into two i.e. the relevant documents and the irrelevant documents. This crisp partitioning are liable to reject relevant items because of strict queries and sometime bring out irrelevant results [19]. Fuzzy modelling IR approach is based on the use of linguistic information at various level in the retrieval process [3]. Another approach employing fuzzy method is for defining flexible query languages capable of capturing the vagueness of user needs as well as simplifying user-system interaction.

Our choice of information retrieval tasks include, but not limited to: Ad hoc retrieval [22], Known item search, and Interactive retrieval [17]. Other variation of this is the classical relevance feedback approach [18].

Designated retrieval activities have formed the basis on which contextual information access, seeking and retrieval is founded. Its importance lies in the fact that, “if you can know your user, you are likely going to treat her/him in a special manner”. We believed and agreed with other proponents of user centred information access that there should be a level of flexibility incorporated into retrieval activities to guide against the problem of “no result in the presence of information”. These inadequacies have been linked to the kind of query and documents representation alongside mode of relevance determination employed by the system.

A fuzzy matching program can operate like a spell checker and spelling-error corrector. For example, a user can types *"Mississippi"* into *Yahoo* or *Google* (both of which employ some level fuzzy matching), a list of hits is returned along with the question, *"Did you mean Mississippi?"* Alternative spellings, and words that sound the same but are spelled differently, are given. A fuzzy matching program can compensate for common input typing errors, as well as errors introduced by optical character recognition (*OCR*) scanning of printed documents. The program can return hits with content that contains a specified base word along with prefixes and suffixes. For example, if *"planet"* is entered as a search word, hits occur for sites containing words such as *"protoplanet"* or *"planetary"*.

A. *The Fuzzy String Matching Model*

The Fuzzy string matching operation is extremely important whenever the search algorithm encounters two strings that are unalike. Our rationale for this task is born out of the fact that whenever there exist no direct relationship between two strings, the strings may still have some things in common. Fuzzy string matching is our attempt to guide against the risk accruable form some class of dirty-data, which include strings that are miss-spelt, inconsistent entries, incomplete context, different ordering and ambiguous data. Consider the strings ‘*onifade’* and ‘*onitade’*. The two strings are practically the same, but for the character‘t’ in the later. The problem arises when a typical matching algorithm encounter this entry, once no direct relationship can be established, it would be ignored. However, when viewed fuzzily, the two strings have a lot in common. Firstly, we can establish that the substring ***‘oni’*** and ***‘ade’*** are in the same position when the two strings are analyzed concurrently. Another point is that they both have the same number of character and thus the main problem is either in misspelling or transposition.

The above described scenario formed the basis for the Fuzzy string matching algorithm analyses shown in figure 3. In other to favourably and concurrently compare the user’s string and the database contents, two dynamic buffers were created at the commencement of the operation. One holds the unmatched characters of the user sub input *‘buffer1’* and the other holds the unmatched characters of the database substring *‘buffer2’*. The algorithm then scans the character content of the two strings concurrently. When the characters are similar, the variable indicating how many characters were matched is incremented. If the characters are dissimilar, the two characters are stored in *buffer1* and *buffer2* respectively. After all the characters might have been compared, it gets to the end of one of the strings (in the case where the size of the two strings are not the same), the fuzzy match value is calculated based on the level of containment or belongingness (via fuzzy membership function ‘func1()) of the *matched character size* and the *size of the database substring* (see Fig. 3). The above operation does not do away with the unmatched characters, instead they are considered to generate some other entries to be displayed alongside the retrieved entries.

While this could generate a high volume of redundant entries, the user has the opportunity to decrease the level of fuzziness and thus reducing the number of entries. We considered the above as exigent for two reasons, extreme cases of misspelling as in the cases of dyslexia, and when the supplied query forms a subset of the database content but not a whole e.g. ‘*Oberman’* and ‘*Hoberman*’.

B. *Unmatched Characters Comparison*

Case study 1: Dyslexia

The word dyslexia is a learning disorder which manifests itself as a difficulty with reading and spelling. It was tagged as a neurological disorder in [14]. It has however been considered in learning disability, language disability, and reading disability [23] etc. Estimates showed that America has between 5 – 17% of her populations in this group of people.

To do this, the unmatched characters placed in buffer1 and buffer2 described above are analysed to check for similarity. The case of dyslexics could be considered as an extreme case, but research has shown that most of “failed-hit” in retrieval operation are due to misspellings. Google and Yahoo search have propose some level of fuzziness to such problems, but the operation is not as robust as what is achieved with Fuzzy String Matching model.

The propensity of human to commit error willingly or otherwise is the basis for ambiguity and lack of desired results in our endeavours, and this factor contributes in no small measure to risk in every facet of our life. We submitted the same input ***‘ONIEADF’*** to Google search engine and the result is interesting. Before discussing the Google result, let us look at the string comparison vis-à-vis the buffering pattern employed by FuzzyMatch to resolve such ambiguity. The model compares the two buffers containing the unmatched characters produced from stage 2 to check for similarity in their character content. The product of this stage is the buffer match value. Once a character presence can be established in the string, the buffer content continues to be manipulated dynamically until the last entry is considered in the string. This results into the fuzzy match which is the multiplicative effect of the buffer match and the level of belongingness. The fuzzy function employed helps to determine the level of fuzziness in the pattern of arrangement of the user’s input and used same to assist in possible rearrangement. The result is shown in figs. 4 below.

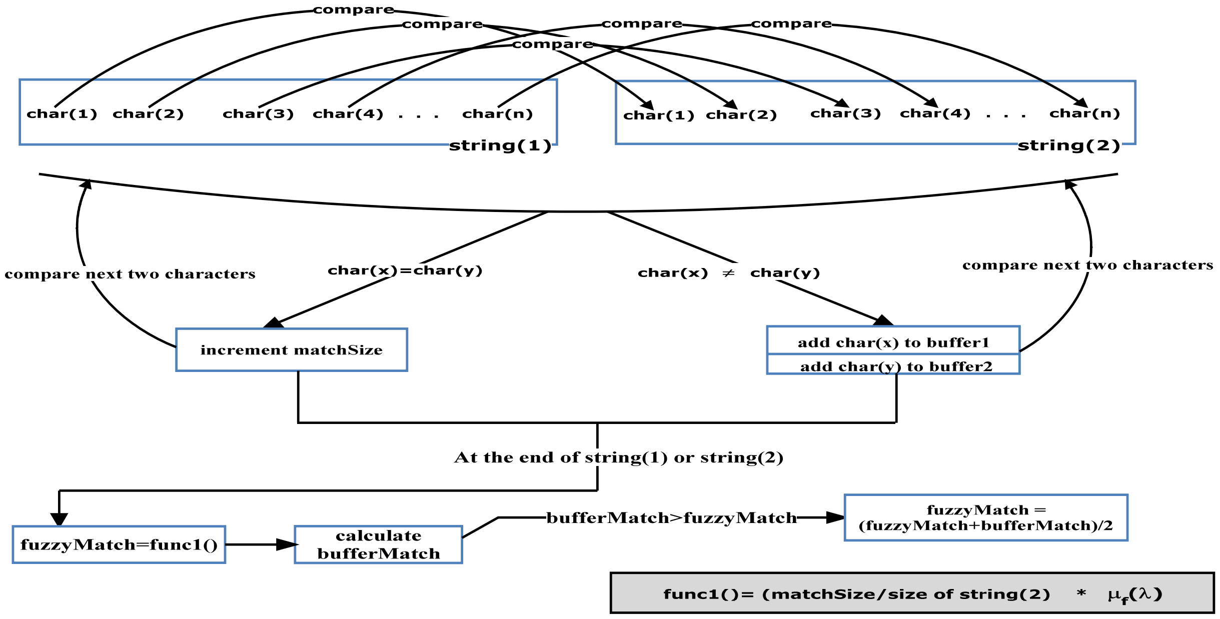

compare
compare
compare
compare
compare
char(1) char(2) char(3) char(4) . . . char(n)
string(1)
char(1) char(2) char(3) char(4) . . . char(n)
string(2)
compare next two characters
char(x)=char(y)
char(x) ≠ char(y)
compare next two characters
increment matchSize
add char(x) to buffer1
add char(y) to buffer2
At the end of string(1) or string(2)
fuzzyMatch=func1()
calculate bufferMatch
bufferMatch>fuzzyMatch
fuzzyMatch = (fuzzyMatch+bufferMatch)/2
func1()= (matchSize/size of string(2) * μf(λ))

Fig. 3: Fuzzy String Matching Model

Common symptoms of a dyslexia include: difficulties in identifying or generating rhyming words, or counting syllables in words (Phonological awareness), difficulties in segmenting words into individual sounds, or blending sound to make words (*Phonemic awareness*). Other include difficulties with word retrieval or naming problem, and difficulties in distinguishing between similar sounds in words, mixing up sounds in multisyllable words (*auditory discrimination*) e.g. “aminal” instead of animal, “bisgetti” for spaghetti. It has however been established that being dyslexic does not inhibit intelligence and thus are equally involved in information search for their information need [10].

Dyslexics for example could spell a word with the same character content but in most cases, the characters are muddled up. For example a dyslexic could spell ‘*clement’* as ‘elcmten’. In order to trap cases like these, the algorithm analyses the character content of the two strings even if their characters do not match concurrently.

String partition for resolving search queries is not uncommon phenomenon, in fact it is interesting to note from our review of search engines operations, more than 80 percent employ string partition. It is therefore not peculiar to our design. Resulting back to our dyslexics query string example of ***‘ONIEADF’,*** we present the same input string to other known search engines and figs. 4 captured the result based on the manner via which our models treats the string and how Google search engine handles the same string.

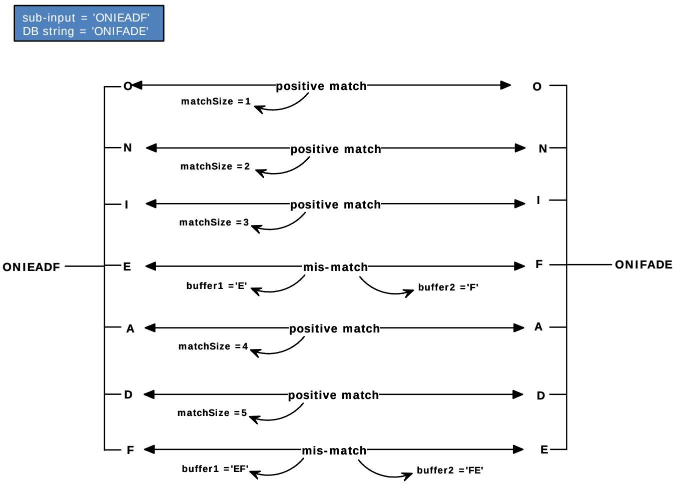
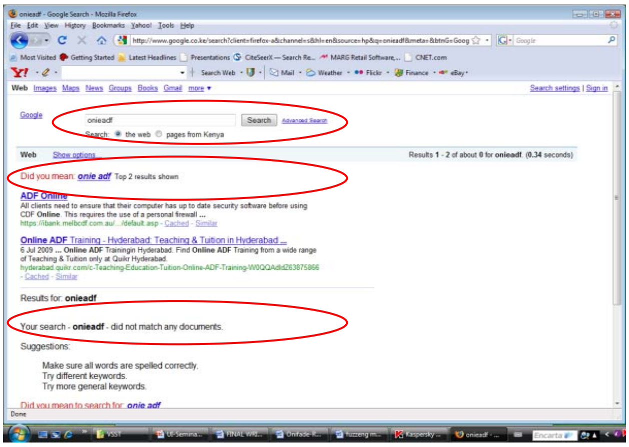


Fig. 4(a): Fuzzy String search result for *'ONIEADF'*

Once the string is typed in by the user, the first thing Google did was to attempt to find a matching pattern in the supplied string. This led to the partitioning of the string to have two other substrings "***onie***" and "***adf***" which is an attempt to see if similar pattern could be established in the database entries. This we say is similar to the string partitioning of our proposed Fuzzy model.


onieadf - Google Search - Mozilla Firefox
File Edit View History Bookmarks Yahoo! Tools Help
Most Visited Getting Started Latest Headlines Presentations CiteSeerX — Search Re... MARG Retail Software,... CNET.com
Search Web Mail Weather Flickr Finance eBay
Web Images Maps News Groups Books Gmail more
Search settings | Sign in
Google
Search
Advanced Search
Web Show options...
Done
VSST UI-Semina... FINAL WRI... Onifade-R... fuzzeng m... Kaspersky ... onieadf - ... Encarta

http://www.google.co.ke/search?client=firefox-a&channel=s&hl=en&source=hp&q=onieadf&meta=&btnG=Goog

onieadf

Search: ● the web ○ pages from Kenya

Results **1** - **2** of about **0** for **onieadf**. (**0.34** seconds)

Did you mean: ***onie adf*** Top 2 results shown

ADF Online
All clients need to ensure that their computer has up to date security software before using CDF **Online**. This requires the use of a personal firewall ...
https://ibank.melbcdf.com.au/.../default.asp - Cached - Similar

Online ADF Training - Hyderabad: Teaching & Tuition in Hyderabad ...
6 Jul 2009 ... **Online ADF** Trainingin Hyderabad. Find **Online ADF** Training from a wide range of Teaching & Tuition only at Quikr Hyderabad.
hyderabad.quikr.com/c-Teaching-Education-Tuition-Online-ADF-Training-W0QQAdIdZ63875866 - Cached - Similar

Results for: **onieadf**

Your search - **onieadf** - did not match any documents.

Suggestions:

Make sure all words are spelled correctly.
Try different keywords.
Try more general keywords.

Did you mean to search for: ***onie adf***

Fig. 4(b): Google search result for *'ONIEADF'*

The two substrings arrived at by Google depict more than the partition visible on the surface. It informs us that presently *Google does not handle any ambiguous substring that is more than four letters presently*; again this is another walk mile away from the functionality of existing search engines. This is evident in the fact that the first three letters of the substring can directly match the desired result however; the supplied fourth character is actually the seventh in the original string which confuses the engine. "***Onie***" therefore has no representation in Google and it has to move on to the next substring for possible resolution. Since character's case (upper or lower) does not affect the result of search any longer, the second substring "***adf***" returns 2 hits which is not in any way close to the desire of the user. As if Google itself knows about the users' dissatisfaction, it declared that "*Your search – 'onieadf' did not match any documents*", and went ahead to give possible suggestions to resolve the problem.

The suggestions from Google are "make sure all words are spelled correctly" – implying that it is recognized that users' input can be wrong and the effect can be adverse to the point of not returning a tangible result. The second is "try different keywords" – this suggestion is only useful if the user knows that the query supplied is wrong. There are occasions when queries are not direct intention of the users, and this is almost always the case because queries are reflection of users' need not the exact. Thus, it might be impossible to supply other keywords on such occasion without deviating from the initial intention. Google also suggests that "try more general keywords" – again this is difficult when you are not very sure where the error comes from.

Case study 2: Misspelling or Query Misrepresentation

This example follows from the analysis by Adam Brookes resulting from the attempted bombing of US flight 253. The excerpt is as follows "*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 example above is very important to tailoring user queries to database entries. In the first case, the saga was described as a "series of human error" which we covered in our work tagged Fuzzontology [15]. The second example follows from the operation of our proposed fuzzy string matching model. With our search engine, even if the name is misspelled, there is going to be tangible retrieval that can be adequately linked to the name as earlier shown above.

V. Conclusion

The success of efficient retrieval operation cannot be over emphasized on the part of users and much more in delivering strategic decisions as in the case of US flight 253. With such error resulting from inadequacies of query reasoning strategy, the world would have lost 290 human beings. It therefore behooves that search tools for the future must be endowed with enough reasoning abilities to resolve ambiguities and uncertainties that engulf query presentation and resolution. In the future, we hope to formally present our prototype of the search tool, and also improve on the reasoning strategy with the inclusion of learning patterns for user's adaptation.

References

[1] BBC News (Adams Brookes) "Obama Announces Security Overhaul" http://newsvote.bbc.co.uk/mpapps

[2] Bordogna G. & Pasi G., Controlling retrieval trough a user-adaptive representation of documents, International Journal of Approximate Reasoning, 12, 317–339, 1995.

[3] Bordogna G. & Pasi G.: Modelling Vagueness in Information Retrieval, in Lectures in Information Retrieval, M. Agosti, F. Crestani and G. Pasi eds., Springer Verlag. 2001.

[4] Broder, A.: "A taxonomy of web search", In ACM SIGIR Forum, Vol. 36, No. 2. (2002), pp. 3-10

[5] Canfora, G. & Cerulo, L.:"A Taxonomy of Information Retrieval Models and Tools" Journal of Computing and Information Technology - CIT 12, 2004, 3, 175–194

[6] Choo, C. W., Detlor, B., & Turnbull, D. Information Seeking on the Web – An integrated model of browsing and searching. *Proceedings of the Annual Meeting of the American Society for Information Science (ASIS)*, pp 1-16, 1999

[7] Crestani F. & Pasi G., Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks, in: “Neuro-fuzzy Techniques for Intelligent Information Systems”, N.Kasabov and Robert Kozma Editors, Physica-Verlag, Springer-Verlag Group, pp. 287–313, 1999.

[8] Crestani, F. & Lalmas, M.: “Logic and Uncertainty in Information Retrieval” In M. Agosti, F. Crestani, and G. Pasi (Eds.): ESSIR 2000, LNCS 1980, pp. 179 - 206, 2000. Springer-Verlag Berlin Heidelberg 2000

[9] Fuhr N, & Buckley C (1991) A probabilistic learning approach for document indexing, ACM Transactions on Information Systems 9:223–248

[10] Morgan, E. & Klein, C.: “The Dyslexic Adult in a non-dyslexic world”. Whurr publishers, ISBN 1861562071, 2000.

[11] Muramatu, J. & Pratt, W.: Transparent queries: Investigating Users’ Mental Models of Search Engines. Proceedings of SIGIR 2001.

[12] Naeza-Yates, R., & Riebeiro-Neto, B.: Modern Information Retrieval, Addison Wesley, NewYork, 1999.

[13] Navarro-Prieto, R., Scaife, M., & Rogers, Y.: Cognitive Strategies in Web Searching. Proceedings of the 5th Conference on Human Factors & the Web, 1999.

[14] O’Toole, Kathleen: “Researchers find white matter defect link to dyslexia” accessed from http://news.stanford.edu/pr/00/000224dyslexia.html

[15] Onifade O.F.W., Thiery O., Osifisan, A.O. & Duffing G. (2010): “FUZZONTOLOGY: Resolving Information Mining Ambiguity in Economic Intelligent Process”. In the Proc. of International Conference on Information Systems, Technology and Management, Bangkok, Thailand, March 11-13. S.K. Prasad et al. (Eds.): ICISTM 2010, CCIS 54, pp. 232–243. Springer-Verlag Berlin Heidelberg 2010

[16] Pasi, G.: “Fuzzy Sets in Information Retrieval: State of the Art and Research Trends”. In H. Bustince, et al., (eds.), *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models,* Springer, pp 517 – 535, 2008

[17] Robins, D.: Interactive Information Retrieval: Context and Basic Notions, Information Science, 3(2), 2000, pp. 57–61.

[18] Rocchio, J.J., Relevance Feedback in Information Retrieval, Prentice Hall, 1971.

[19] Salton G., Automatic Text Processing - The Transformation, Analysis and Retrieval of Information by Computer, Addison Wesley Publishing Company, 1989.

[20] Smeaton, A.F. Progress in the application of Natural Language Processing to Information Retrieval tasks. The Computer Journal, 35(3):268-278, 1992.

[21] van Rijsbergen, C. J. Information Retrieval. London: Butterworth’s, 1979.

[22] Voorhees, E.M & Harman, D.K (2000). *Overview of the Eighth Text Retrieval Conference* (TREC-8). In: Information Technology: The Eighth Text Retrieval Conference (TREC-8). NIST SP 500-246, pp.1-23, GPO: Washington, D.C

[23] Wiki, 2010 “Dyslexia” http://en.wikipedia.org/wiki/Dyslexia. 08 January, 2010.

[24] Witel, G. L & Wu, S.F. (2004): “On attacking statistical spam filter”. CEAS: First Conference on Email Anti-Spam, 2004