Journal of Computer and Mathematical Sciences, *Vol.9(10),1556-1567 October 2018* *ISSN 0976-5727 (Print)*
(An International Research Journal), *www.compmath-journal.org* *ISSN 2319-8133 (Online)*

An Efficient Indexing Mesh Term Description Logic Using in Medical Subject Headings

R. Aravazhi[1*] and M. Chidambaram[2]

[1]Ph.D., Department of Research Scholar, Computer Science,
A.V.V.M Sri Pushpam College (Autonomous),
Poondi, Thanjavur, Tamilnadu, INDIA.
[2]Assistant Professor in Computer Science,
Rajah Serfoji Government College (Autonomous),
Thanjavur, Tamilnadu, INDIA.
email: [1]aravazhi.r@gmail.com, [2]chidsuba@gmail.com.

(Received on: October 10, 2018)

ABSTRACT

In this research, propose another methodology for ordering biomedical records in light of a possibilistic organize that completes halfway coordinating among archives and biomedical vocabulary. The principle commitment of our methodology is to manage the imprecision and vulnerability of the ordering undertaking utilizing probability hypothesis. The biomedical records indexing is a genuinely touchy stage in the data recovery. Be that as it may, terms introduced in an archive are not adequate to totally speak to it. At that point, the abuse of the certain data, through outside assets, is essential for better ordering. For this research, another indexing mesh term description logic (IMTDL) model for biomedical archives in view of depiction rationales has been proposed to produce pertinent files. The records and the outer asset are spoken to by engaging articulations; a first measurable stage comprises in doling out a significance degree to each term in the report and a semantic part to remove the most critical ideas of the MESH thesaurus (Medical Subject Headings). The idea extraction step utilizes the portrayal rationales to consolidate the factual and semantic methodologies pursued by a cleaning part to choose the most essential lists for the record portrayal. For the tests stage we utilized the OHSUMED accumulation, which demonstrated the viability of the proposed approach and the significance of utilizing depiction rationales for the ordering procedure.

Keywords: MESH, Indexing, Description logic.

1. INTRODUCTION

To enhance the execution of a data recovery framework, it is basic to build up a programmed ordering framework that can have as yield the most delegate file of a record. The last might be spoken to by free or controlled terms. Controlled vocabulary might be a basic arrangement of terms or an organized arrangement of terms (or ideas) connected by various leveled or affiliated relations (synonymy, more extensive than, smaller than, et cetera, for example, the Medical Subject Headings® (MeSH) thesaurus. Utilizing the semantic properties of a semantic asset (thesaurus, wording, cosmology, etc.), an applicable term (or idea) might be separated despite the fact that it doesn't happen in the archive. On account of free term ordering, catchphrases are removed without utilizing any information asset. Consequently, the list isn't known previously and extricated terms may not adjust to the subject to which the record has a place. Also, applicable terms that don't happen in the archive can't be separated.

The data recovery differs from a few gatherings for the most part ordering procedure, the last not comprises just to characterize reports but rather likewise ordering them. Human ordering is bounty costly and ask for a considerable measure of time, so analysts plan of action to utilize programmed ordering framework to take care of a piece of the issue. A few methodologies have been proposed for the programmed ordering of reports. View to the huge measure of data exists on the web and with the expanding of clients requests for important data, the advancement of pertinent data recovery frameworks is important to address these issues. They can be grouped into two primary families a free ordering and a controlled ordering, the principal comprises to speak to the reports just by the terms present in the record, there is no information of the space and the ordering terms are not arranged in advance. The second family utilizes outside assets for the portrayal of a record, for example, philosophy, thesaurus and scientific classification.

Different components are powerful in the ordering procedure, the rough coordinating can enhance the reports ordering, it permits to discover in the record different variations of the controlled vocabulary to speak to it and furthermore makes it conceivable to remove the compound terms that offer an arrangement of words with the archive, for this situation the decrease of words to their foundations (stemming) has an incredible significance for better ordering. Then again, correct coordinating makes it conceivable to discover in the record just the ideas that exist in the vocabulary. Jumpers rationales have been utilized to enhance the execution of the ordering frameworks, for example, Fuzzy rationale, which is an augmentation of established rationale that permits the displaying of blemishes information and approximates in some proportion of the human thinking adaptability, the other rationale in light of Fuzzy sets hypothesis is the possibilistic rationale and the probabilistic rationale. As far as anyone is concerned, the description logic are not utilized in the biomedical ordering process. At that point, so as to appraise the importance of an idea to a report, proposed an ordering approach in view of the portrayal rationales, the description logic are utilized to speak to phrased learning of an application field in an organized and formal way, they are extremely successful and effective and have been connected in different areas, however very little in the data recovery field. These rationales make it conceivable to separate understood informations, from an outside asset, which can speak to an archive yet they are absent in this last mentioned.

2. RELATED WORK

A few methodologies have been proposed for the ordering of biomedical unstructured reports: measurable methodologies, semantic techniques and others that join free ordering and controlled ordering. In its factual methodology, utilizes the Conditional Random Fields demonstrate for catchphrase extraction. After the preprocessing stage and after the report division, each word is given a level of significance relying upon its position utilizing the Conditional Random Fields; the other advance comprises in extricating the most imperative words by contrasting them and the manual task. Propose an ordering model in view of factual and etymological information, in the semantic setting they utilize the CRF for expressed extraction in light of the perception of the phonetic attributes of the significant terms and their neighbors in the content and for the separating procedure, they depend on measurable learning of the content with a specific end goal to give more significance to the applicable terms[4].

Another measurable technique for the free extraction of terms, Propose an ordering model which utilizes a report not all the corpus, the most continuous terms are separated in the primary stage and after the age of terms which rehashes in a similar sentence pursued by a weight task stage to extricate the most vital catchphrases. The other family is controlled ordering, in this family the creators depend on outer asset for the terms extraction. In her methodology, utilizes possibilistic rationale to coordinate the archive terms somewhat to the thesaurus MeSH (Medical Subject Headings) terms. The fundamental part in this methodology is the treatment of imprecision and vulnerability. To enhance the estimation of report pertinence given an idea, she utilizes two measures: the likelihood and the need. The likelihood considers the level of dismissal of an unimportant report given an idea. Need estimates how well a report is applicable to the idea. A change part decreases the points of confinement of fractional coordinating which produces superfluous data, despite the fact that this sort of coordinating makes it conceivable to discover in the archive different variations of the controlled vocabulary. A third stage committed to sifting, which utilizes information given by UMLS (Unified Medical Language System) to keep those applicable ideas. The other semantic methodology, utilize in excess of twenty biomedical ontologies from the National Center for Biomedical Ontology, this device utilize the terms of these ontologies to clarify and recognize naturally the literary depictions that exist in the distinctive assets, and register a score for every comment created by class (favored term, non-favored term or equivalent word). Built up an ordering model in light of Bayesian systems utilizing the structure of an outer asset, just progression and connections between descriptors are abused in the thesaurus, which are a decent preferred standpoint to utilize some other thesaurus[2]. This creator has made a Bayesian system where there are diverse attributes of the thesaurus (proportionality connections, pecking order), and a short time later given a record to characterize, its terms are instantiated in the system and a probabilistic surmising calculation processes the back likelihood in the thesaurus.

The third family is what consolidates free ordering and controlled ordering, in this family we discover half breed ordering models that depends on factual figuring or phonetic handling pursued by abuse of an outside asset to extricate the most important terms for an archive ordering. In her methodology BioDI (A New Approach to Improve Biomedical Documents Indexing) in light of VSM (Vector Space Model), joined two strategies factual and semantic techniques for the descriptors extraction which can speak to a record, the main stage comprises to set up all terms in corpus and thesaurus by coordinating the stemming system to uncover every morphological variation of terms. For the factual methodology, this creator utilized the tf-idf measure to figure the significance of each word in the report and utilized a similar measure for the semantic technique to process the significance level of each term in the MESH thesaurus. The archive is spoken to by a vector and the term of the outer asset also, after, the likeness between a record and a term is processed utilizing cosine closeness. When this stage is finished, the score of a descriptor is figured by concurring the scores of its terms where it will take the most extreme score of these. The sifting step is to let the most critical descriptors by abusing the learning of UMLS (Unified Medical Language System).

The first is a measurable technique which comprises in extricating the references words from the normal dialect thinking about the compound words. For the semantic technique, the MESH thesaurus takes as info the words as of now removed in the primary stage to uncover the arrangement of terms, and afterward the thesaurus is spoken to as a record where each term is spoken to by the references words[7]. After these, terms will be requested by connecting the references words to make a sentence; a term is perceived if every one of its words are available in the sentence by coordinating the division of each sentence of the content. Proposed a methodology called MaxMatcher in light of a factual figuring for the extraction of terms which are approved then by a query lexicon, the methodology comprises in choosing the most huge expressions of the record to a given idea, the idea will be chosen in the event that one of the words that shape it has the better score. The last record contains all the most important ideas utilizing the UMLS Meta thesaurus design. Factual methodologies abuse the archive terms factual measures i.e. just these terms can speak to the archive, furthermore alternate methodologies abuse, outside assets with correct coordinating which diminishes exactness, different strategies less important and different rationales. The proposed approach consolidates two strategies: measurable and semantic, the abuse of LDs in the ordering procedure permits a superior portrayal of a report and Mesh terms.

3. METHODOLOGY

Description Logic are a group of dialects of learning portrayal that can be utilized to speak to the expressed information of an application space in a formal and organized way. The portrayal rationales name alludes, from one viewpoint, to the depiction of ideas used to portray an area and, then again, to the semantics in light of rationale which can be given by a translation to predicates rationale of the primary request. The primary target of Description Logic is to have the capacity to reason successfully to limit reaction time. Thus, mainstream researchers has distributed numerous examines on the connection among expressiveness and execution of various Description Logic. The fundamental nature of Description Logic lies in their deduction calculations, whose many-sided quality is frequently substandard compared to the complexities given by first-arrange rationale. The expressed level: portrays the general information of a space and how they are interrelated, a Terminological Box incorporates meaning of ideas and jobs. The truthful level: portrays a particular or a nearby data, an Assertional Box contains an arrangement of statements about people: having a place affirmations and job attestations. The deduction is completed at expressed or genuine level. In the expressed level, four induction issues emerge, satisfiability, comparability, subsumption, and disjunction. A similar number of derivation issues in the verifiable level: Consistency, recuperation issue, job checking and occasion checking.

1) Preparation stage
2) Documents Representation / MESH terms representation with description logics
3) Extracting concepts using description logic
4) Filtering to keep relevant concepts extracted in the previous step.

Preparing documents and MESH terms. This step consists of preparing terms and words for the next step, it can be called the pretreatment phase. Each document or term goes through the following tasks: Divide document into sentences, Remove empty words, punctuations, symbols and numbers, Separate sentences into words, Stemming: is the technique that allows, transforming a word into root, a root corresponds to the part of the remaining word after the removal of its suffix,

D: Set of documents in the collection, ND: Number of documents in the collection, NM: Mesh concepts Number, i,j: Counters.

```
Result: List of Indexing concepts
Initialize
Term list ;
for i=1 to ND do
        Term list = Term list + pretreated Term$_i$
end
Record Descriptive Expression;
for i=1 to ND do
        Term weight$_i$   word weight in the document
(function (1))
if Term weighti _ threshold then
        Record Descriptive Expression=   Record Descriptive Expression + Wordi
end
end
Mesh Descriptive Expression (MeshDE) ;
        for j=1 to NM do
MeshDE   MeshDE + pretreated MeshTermj
end
/* Concepts extraction */
```

```
Candidate Concepts (CC)   ;
for i=1 to ND do
for j=1 to NM do
        CC + $concept_j$ extracted with the
inference algorithm
end
end

Primary Index PI   concepts that all the words of its chosen term are present in the document
Secondary Index SI   concepts that some words of its terms are present in the document
if ((concept C1 € AI)2 the same Mesh descriptor of
(concept C2 € NI)) then
NI NI+C1
end
Return (List of Indexing concepts NI)
End
```

Words of the same root are grouped in the same class. The stemming process makes it possible to group the different forms of the same words that is having the same semantic, which makes it possible to find other variants of controlled vocabulary in the document and optimize the indexing process.

This part is devoted to present the document words and the MESH terms by the description logics, a statistical compute is made to give an importance degree to each document word, after, this document will be represented by a descriptive expression and the same process will be applied to MESH terms.

Medical record document representation with DLs: After the word preparation phase, a score is assigned to each word in the document with BM25 measure. The weight of a word in a document is calculated with the following equation (1):

$$Ww_j = WFDoc_j * X$$

$$Where$$

$$X = (\frac{\log(\frac{N - n_j + 0.5}{n_j + 0.5})}{WFDoc_j + k_1 * ((1 - b) + b\frac{DL}{AvDL})}) \qquad (1)$$

After words weighting phase, the next step is to represent the document with a descriptive expression. Each document term (termj) represents an indexing element (Index element$_j$) and all of these elements form a descriptive expression, the latter represents this document (D$_i$). The final index (Index doc) contains the set of descriptive expressions since each document represented by an expression.

The descriptive expression is of the following form:
Term$_j$ $\equiv$ *Index elemen*$_j$
D$_i$ $\equiv$ *denoted by:Index element*$_1$
denoted by:Index element$_2$
denoted by:Index element$_3$
denoted by:Index element$_n$
Example
D$_i$ $\equiv$ *denoted by:Acetaldehyde*
denoted by:Buffers
denoted by:Catalysis
denoted by:P ancreatic

Work is a reference thesaurus in the biomedical field, it is a controlled vocabulary base, this thesaurus incorporates a rundown of terms having progressive, synonymic and closeness relations between them. A MESH idea is spoken to by a favored term and others not favored, an idea shapes a spellbinding articulation of the accompanying structure:

Con$_i$ $\equiv$ *PT described_by.T*$_1$ *described by:T*$_2$
Described_by.T$_3$ *.. described_by.T*$_n$
Example
AbdoninalInhuries $\equiv$ *PT described_by.AbdoninalInhuries*
Described_by:AbdominalInjury
Described_by:Injuries, Abdominal

The information base is comprised by ideas (Con,PT,T) and jobs (portrayed by), Con is an idea of MESH thesaurus, PT present the favored term and T is a non-favored term. This articulation is a combination of something like one term that serves to recognize the MESH idea, it is conceivable to contain other non-favored terms to refine the semantic depiction of the idea, portrayed by speaks to the job that connections these ideas. The arrangement of these articulations frame the MESH thesaurus list (Index MESH).

To extricate the most delegate ideas for an archive, thinking is performed at the phrased level (TBox). To play out this errand, we utilized a coordinating in light of the count of subsumption (vS). In this unique circumstance, an idea C1 is subsumed by an idea C2 for a wording T, if and just if CI1 v CI2 for all model I of T.

Hypothetical level beneath and clear level above. The hypothetical level incorporates an arrangement of archives and an arrangement of MESH terms as spoke to in the ordinary express, the spellbinding level speaks to the arrangement of reports in a pack (Index doc) and the arrangement of MESH terms in another sack (Index MESH) utilizing elucidating articulations. The ideas extraction is finished by the coordinating between a report in Index doc and an idea in Index MESH, this coordinating must check Con$_i$ versus Doc$_i$ for our insight base. At last, the arrangement of hopeful ideas for an archive.

The objective of this progression is to keep just the appropriate ideas, among the applicant ideas, that can record an archive. For this reason, the extricated ideas in the past segment have been separated into two packs: the first is the Primary index (PI) which incorporates ideas that every one of the expressions of its favored term are available in the record, the secondary index (SI) contains the arrangement of ideas which a few expressions of its terms (favored and non-favored) are available in the archive. To include an idea C from PT to the SI, we misuses the thesaurus MESH design, at that point it is important to know whether C has a connection with the Necessary list ideas. On the off chance that the idea C has a place with a similar descriptor of one of the Necessary list ideas, C is added to the Necessary list. The last stage permits grouping the ideas as indicated by their scores, the score of an idea is the whole of the weights of its words present in the report. The heaviness of a word is a similar weight figured in the measurable part. The primary n ideas will be chosen to speak to the record last file and in this manner contrasted and human ordering.

4. EXPERIMENTAL RESULT

To test the execution of the proposed approach, we utilized a sub-corpus of the OHSUMED gathering made out of 5000 archives in english, each report is in a reference frame comprising of title, unique, MeSH ordering terms, creator, source, and distribution compose. We are intrigued just by the title and the unique since they contain the data that we requirement for our experimentation. These two sections have a similar level of significance. We utilized additionally another corpus comprising of titles and modified works of 2000 haphazardly chose CISMeF (Catalog and Index of Medical Sites in French) assets. Three sorts of reports are recorded in CISMeF: archives for patients, suggestions and reports for instructing. In all trials we kept just the initial fifteen ideas in the last file. Truth be told, the normal number of ideas in manual records in MEDLINE is 15.

Table 1: Test Records

	OHSUMED	CISMeF
Total number of records	7000	2000
Average words number in heading	10.7	10.2
Average words number in summaries	110.5	105.1
No of records for patients		700
Number of documents of Recommendation Type	—	450

To evaluate the proposed approach, we used three standard measures intended to evaluate the obtained answers in relation to the user expectation, precision (P), recall (R) and f-measure (Fm). Precision is the ratio between the number of correct concepts and the total number of extracted concepts. The recall is the ratio between the number of correct concepts and the number of concepts that corresponds to manual indexing. Fmeasure combines precision and recall, is the harmonic mean of precision and recall. The concepts are considered correct if they belong to the manual index. These three measures are calculated with the following formulas:

$$\text{P}=\frac{TNCT}{TNTAG} \tag{2}$$

Where

NCC is the Total number of correct terms
TNCAG is Total number of Terms Automatically Generated

$$\text{P}=\frac{TNCT}{NTME} \tag{3}$$

Where
NTME Number of Term Manually Extracted

$$Fm = 2 * \frac{(P * R)}{(P + R)} \tag{4}$$

The point of these trials is to compute the estimations of the distinctive estimates that were refered to previously, review, exactness and f-measure. We did this examination utilizing the two corpus OHSUMED and CISMeF. The tests on the two corpus produced nearly similar outcomes so we introduced just the outcomes that worry a solitary corpus (OHSUMED). We contrasted the proposed approach and different methodologies, BioDI and Yutaka approach, in light of controlled vocabulary and free terms, to see the distinction between the abuse of controlled vocabulary and the utilization of free terms, and in addition the distinction for the learning portrayal between the Vector space show and the DLs. To see the significance of IMTDL approach, the five after investigations: the primary, the 3 first, the 5 first, the 10 first and the 15 first ideas removed.

We indicate that the normal number of ideas picked in manual ordering for MEDLINE references is 15. By breaking down the accuracy bend, plainly the proposed approach is superior to other people. Superfluous or unimportant ideas comprise what is called commotion. Exactness is against this clamor, and from the outcomes got it tends to be said that a couple of insignificant ideas are proposed, contrasted and BioDi and Yutaka and that indexing mesh term description logic (IMTDL) can be considered as “exact”. For the recall measure, the three methodologies results are close and have mean qualities, and after that the quantity of displayed ideas is normal. Then again, on the off chance that we have many intriguing ideas however these are not in the appropriate responses list, we are discussing quietness. Quietness is against the review. The proportion between the exactness and review measures is appeared in the above bend, it appears to be evident that the IMTDL approach has preferred execution over others. Indeed, even ideas that don't have their words in the record might be significant. Starting here, the utility of utilizing outer assets and misuse of incomplete coordinating strategy.

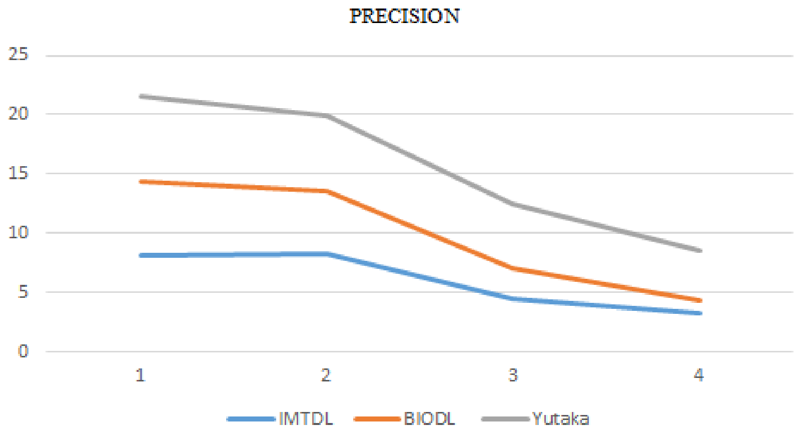

Fig 1. Precision Result

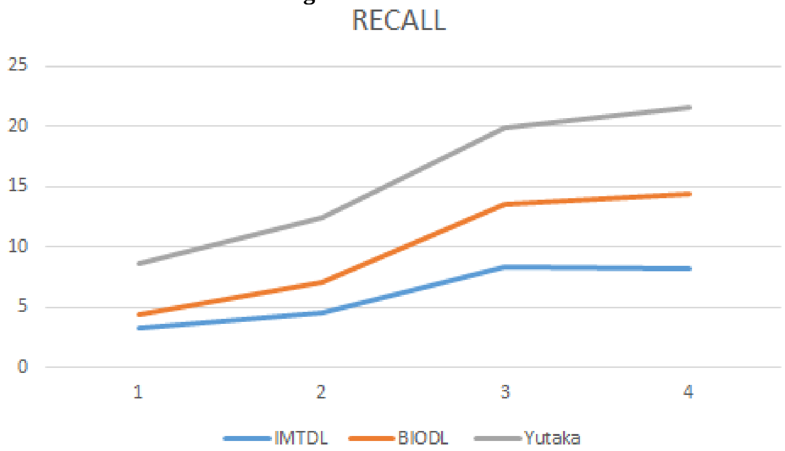

Fig 2. Recall Result

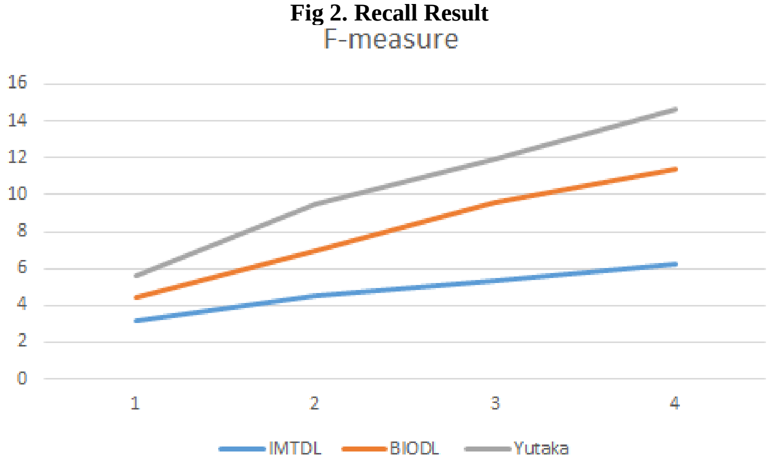

Fig 3. F-measure Result

5. CONCLUSION

In this research, reformulate reports and outer assets as far as depiction rationales by description logic algorithm. It utilize the depiction rationales to join the measurable and semantic calculations pursued by a cleaning part to choose the most vital records for the report portrayal. Proposed approach demonstrates awesome effectiveness particularly with the utilization of fractional coordinating and stemming strategy. The trial indicates obviously the enthusiasm of the proposed calculation and particularly with the utilization of description. The later have greater expressivity and more execution for the information portrayal, and additionally the nature of the deductions calculations. As a future work, intend to apply the proposed way to deal with other standard corpus, for example, CISMeF, the consequences of the last will be distributed in future work, and utilize another word reference in the sifting stage to uncover more relations between ideas. We likewise plan to contemplate the multifaceted nature of the proposed algorithm and to think about it versus the other man approaches proposed in the writing.

REFERENCES

1. A.Elbehi, M. N. Omri, and M. A. Mahjoub, “Possibilistic reasoning effects on hidden markov models effectiveness,” *The 2015 IEEE International Conference on Fuzzy Systems*, vol. 33, pp. 1–8, (2015).
2. W. Chebil, L. F. Soualmia, M. N. Omri, and S. J. Darmoni, “Indexing biomedical documents with a possibilistic network,” *Journal of the Association for Information Science and Technology*, vol. 67, pp. 928–941, (2016).
3. L. A. Zadeh, “Fuzzy sets as a basis for a theory of possibility,” *Fuzzy Sets Syst*., vol. 100, pp. 9–34, (1999).
4. S. Radhouani, G. Falquet, and J. Chevallet, “Description logic to model a domain specific information retrieval system,” vol. 5181, pp. 142–149, (2008).
5. S. Radhouani and G. Falquet, “Description logics-based modelling for precise information retrieval,” pp. 1–11, (2008).
6. F. Fkih and M. N. Omri, “Complex terminology extraction model from unstructured web text based linguistic and statistical knowledge.” *IJIRR*, vol. 2, pp. 1–18, (2012).
7. C. Jonqueta, P. LePendua, S. Falconera, A. Couleta, N. F. Noya, M. A. Musena, and N. H. Shaha, “Ncbo resource index: Ontology-based search and mining of biomedical resources,” Semantic Web Challenge, 9th International Semantic Web Conference, ISWC’10, pp. 316–324, (2010).
8. L. M. de Campos, J. M. Fern´andez-Luna, J. F. Huete, and A. E. Romero, “Automatic indexing from a thesaurus using bayesian networks: Application to the classification of parliamentary initiatives,” *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 865–877, (2007).
9. A.Happe, B. Pouliquen, A. Burgun, M. Cuggia, and P. L. Beux, “Automatic concept extraction from spoken medical reports.” *I. J. Medical Informatics*, vol.70, pp.255-263, (2003).

10. X. Zhou, X. Zhang, and X. Hu, "Maxmatcher: Biological concept extraction using approximate dictionary lookup," *International Conference on Artificial Intelligence*, pp. 1145–1149, (2006).
11. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds., The Description Logic Handbook: Theory, Implementation, and Applications. New York, NY, USA: Cambridge University Press, (2003).
12. V. Haarslev and R. Mller, "Description of the racer system and its applications," The International Workshop on Description Logics, pp. 132–141, (2001).
13. B. Parsia and E. Sirin, "Pellet: An owl dl reasoner," In Proceedings of the International Workshop on Description Logics, p. 2003, (2004).
14. D. Tsarkov and I. Horrocks, "Efficient reasoning with range and domain constraints," pp. 41–50, (2004).
15. K. Boukhari and M. N. Omri, "RAID: Robust Algorithm for stemming text Document," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 8, pp. 235–246, (2016).
16. K. boukhari and M. N. Omri, "Said: A new stemmer algorithm to indexing unstructured document," The International Conference on Intelligent Systems Design and Applications, pp. 59–63, (2015).
17. S. E. Robertson and S. K. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, vol. 27, pp. 129–146, (1976).