DOI: 10.15439/2015F130
ACSIS, Vol. 5

The data retrieval optimization from the perspective of evidence-based medicine

Vladimir Dobrynin*, Julia Balykina*, Michael Kamalov*,
Alexey Kolbin†, Elena Verbitskaya† and Munira Kasimova‡
*Saint-Petersburg State University, Universitetskaya nab., 7-9, Saint-Petersburg, Russia
Email: v.dobrynin@bk.ru, Email: {julia.balykina, mkamalovv} @gmail.com
†Pavlov First Saint-Petersburg State Medical University, L'va Tolstogo str., 6/8, Saint-Petersburg, Russia
Email: alex.kolbin@mail.ru, Email: elena.verbitskaya@gmail.com
‡Tashkent Institute of Postgraduate Medical Education, Parkent str., 51, Tashkent, Uzbekistan
Email: drkasimovamunira@mail.ru

***Abstract*—The paper is devoted to classification of MEDLINE abstracts into categories that correspond to types of medical interventions - types of patient treatments. This set of categories was extracted from Clinicaltrials.gov web site. Few classification algorithms were tested including Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM implementations from sklearn machine learning library. Document marking was based on the consideration of abstracts containing links to the Clinicaltrials.gov Web site. As the result of an automatical marking 3534 abstracts were marked for training and testing the set of algorithms metioned above. Best result of multinomial classification was achieved by Linear SVM with macro evaluation precision 70.06%, recall 55.62% and F-measure 62.01%, and micro evaluation precision 64.91%, recall 79.13% and F-measure 71.32%.**

I. Introduction

At THE moment an evidence-based medicine approach is actively developing in medical practice. This approach requires an expert to choose a method of patient treatment based on available evidences of safety and efficiency of the method. Complexity of evidence-based medicine application in practice involves not only control over saving new research results, but also assessment of the quality and reliability of existing ones. To solve this problem in evidence-based medicine a grading scale for ranking studies by level of evidence is used. For example, in the USA the National Guideline Clearinghouse[1] recommends to follow levels of evidence and grades (table I, table II).

TABLE I
Levels of Evidence

I A	Evidence from meta-analysis of randomized controlled trials (RCTs)
I B	Evidence from at least one randomized controlled trial
II A	Evidence from at least one controlled study without randomization
II B	Evidence from at least one other type of quasi-experimental study
III	Evidence from non-experimental descriptive studies, such as comparative studies, correlation studies, and case-control studies
IV	Evidence from expert committee reports or opinions or clinical experience of respected authorities, or both

TABLE II
Grades of Recommendations

A	Directly based on Level I evidence
B	Directly based on Level II evidence or extrapolated recommendations from Level I evidence
C	Directly based on Level III evidence or extrapolated recommendations from Level I or II evidence
D	Directly based on Level IV evidence or extrapolated recommendations from Level I, II, or III evidence

However, in some cases studies corresponding to the first level of evidence may contain errors in the correctness of randomized controlled trials (RCTs). More detailed description of some examples with errors in studies with the first level of evidence is reviewed in [1]. Solution of the problem is the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system. This system evaluates level of evidence for different studies and ranks them by recommendation significance with due consideration of additional criteria for evaluation. Additional criteria for GRADE are presented in table III and described in more detail in [2]. GRADE considers only two classes of recommendations: strong or low-level recommendations. The quality level of evidence is presented in 4 levels. Thus, using additional GRADE factors it becomes possible to rise or lower the value of research.

Another actively developing research trend is information retrieval application in the field of medicine based directly on the use of MEDLINE[2] database. For example, in [3] MEDIE search engine developed for MEDLINE database, that executes semantic search, keyword search and generalized concordance lists (GCL) search is described. In [4] the Hierarchical Hidden

[1] http://www.guideline.gov/

[2] http://www.nlm.nih.gov/

978-83-60810-66-8/$25.00©2015, IEEE

TABLE III
QUALITY ASSESSMENT CRITERIA

Study design	**Quality of evidence**	**Lower if**	**Higher if**
Randomized trial	High	Risk of bias -1 Serious -2 Very serious Inconsistency -1 Serious -2 Very serious Indirectness -1 Serious -2 Very serious Imprecision -1 Serious -2 Very serious Publication bias -1 Serious -2 Very serious	Large effect +1 Large +2 Very large Dose response +1 Evidence of gradient All plausible can founding +1 Would reduce a demonstrated effect or +1 Would suggest a spurious effect when results show no effect
	Moderate		
Observational study	Low		
	Very low		

Markov Models algorithm for retrieving information about protein and its location from the MEDLINE abstract database is considered. Study [5] considers a method of automatic term extraction developed specifically for indexing documents from large medical collections. Computational experiments are conducted on a set of documents from MEDLINE database. In [6] an unsupervised clustering technique called SOPHIA is presented, that is evaluated on the MEDLINE testing set collection. Study [7] describes an experiment that changes the ranking strategy using the term-graph data structure for assessing the importance of a document to a user's query to the MEDLINE database. In [8] existent question-answering system based on principles of evidence based medicine is presented. Study [11] describes a fuzzy VIKOR framework for ranking internet health information providers.

Based on the relevance and demand for joint studies in the field of medicine and information retrieval, it was decided to start a development of a search engine for the MEDLINE database on the basis of the Saint-Petersburg State University with the support of Pavlov First Saint-Petersburg State Medical University and Tashkent Institute of Postgraduate Medical Education. The main goal of the project is to develop a new ranking method for search results, which takes account for a level of evidence and GRADE criteria.

II. PROBLEM STATEMENT

The object of analysis of this work are documents containing abstracts of articles from the MEDLINE international database of medical research. The goal is to group abstracts according to subtypes of medical interventions. Subtypes of medical interventions correspond to various methods for patient treatment and prophylaxis. Examples of medical intervention subtypes were taken from the Clinicaltrials.gov[3] Internet resource:

1). Drug;

2). Biological;

3). Device;

4). Dietary Supplement;

5). Procedure;

6). Radiation;

7). Behavioral;

8). Genetic;

9). Other.

This website is a public register approved by the US International Committee of Medical Journal Editors. It provides relevant structured information about conducting clinical studies for a wide range of diseases.

The assigned task falls within the domain of machine learning and requires an implementation of the following auxiliary problems:

1). Development of a method for automatic markup of abstracts from a training and test set by medical intervention subtypes, based on the existence of a link between documents that represent paper abstracts from MEDLINE database and contents of registered clinical trials on Clinicaltrials.gov. The link is presented by reference to Clinicaltrials.gov Web resource.

2). Training methods for multinomial classification by means of selected set of classical algorithms such as Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM from the sklearn[4] library for further evaluation and selection of a more effective algorithm.

First it was decided to evaluate linear multinomial classification algorithms for an obtained marked sample of MEDLINE abstracts. Therefore, the following linear algorithms were chosen: Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM. In the future it is planned to choose a set of nonlinear multinomial classification algorithms and conduct experiments with the same marked sample of MEDLINE abstracts.

III. METHODS FOR ABSTRACTS MARKUP

At the first stage of handling the problem 90 paper abstracts of the year 2011 taken from MEDLINE database and

[3] https://clinicaltrials.gov/

[4] http://scikit-learn.org/

which contained links to the Clinicaltrials.gov Web site were examined. To simplify abstract processing, an abstracts.xml document was created which has the following structure (per entry). Document structure (1):

```
<document>
   <doc_id></doc_id>
   <date></date>
   <title></title>
   <body></body>
   <topics></topics>
   <place></place>
   <author></author>
   <type></type>
</document>
```

where: <document> is a document container; <doc_id> is a paper identifier; <date> – publication date of the article; <title> – title of the article; <body> – body of the abstract; <topics> – article keywords; <place> – journal where the article was published; <author> – paper authors; <type> – a subtype of medical intervention. All abstracts were transfered into this structure. Every record in abstracts.xml satisfies the structure listed above.

These 90 documents were marked manually based on the search for links between abstracts and Clinicaltrials.gov Web resource. Linkage was performed by finding in abstract a reference (eg., NCT00893711). Such a reference meant that the study was indexed at Clinicaltrials.gov. The sequence of manual markup was as follows:

1) A link to Clinicaltrials.gov site was retrieved from an abstract.

2) With the help of Clinicaltrials.gov internal search engine a search was performed in order to find studies corresponding to the reference given in the abstract. An example of using internal search engine is presented in Fig. 1.

3) After the search, a found subtype of medical intervention was manually added to the document structure in the <type> </type> field. It was proposed to impose a restriction on the Clinicaltrials.gov web-resource search results: if the study was represented by two subtypes of medical intervention as it is shown in Fig. 1, it was suggested to use the first subtype because it contains the main information about the study.

As a result of manual marking it was possible to group 60 out of 90 abstracts into the following subtypes: Behavioral, Biological, Device, Dietary_Supplement, Drug, Other, and Procedure. The remaining 30 abstracts were divided into 4 groups:

1) no link to clinicaltrials.gov;

2) contains a link, but no information about the subtype of medical intervention;

3) contains a link but studies are of observational type;

4) contains a link with an error (eg., ISRCTN51481987).

It was further decided to consider corpus containing 2000000 abstracts for years 2006 to 2013 and automate the process of markup. An automation has been implemented with the help of an application developed in python that performs a search for links in abstracts (eg. NCT00893711) as a regular expression and a web-crawler that searches through links http://clinicaltrials.gov/show/NCT00776256?resultsxml=true replacing the part (NCT00893711) with the one found in the abstract and extracting data from the xml page contained in the <intervention_type> field. Results extracted from <intervention_type> field were then added to the <type> </type> field of document structure (1) using the developed software application. The result of parsing xml pages also imposes restrictions: in case of two fields <intervention_type> is marked by the first field.

As a result of an automatic marking of 2000000 abstracts, 3534 abstracts were marked. Remaining abstracts were divided into groups:

1) no link to clinicaltrials.gov;

2) contained a link, but when referring to web-crawler on corresponding page an error appeared "404 - page not found";

3) contained a link but corresponding studies were of observational type;

4) contained a link with an error (eg., CTNO1481987).

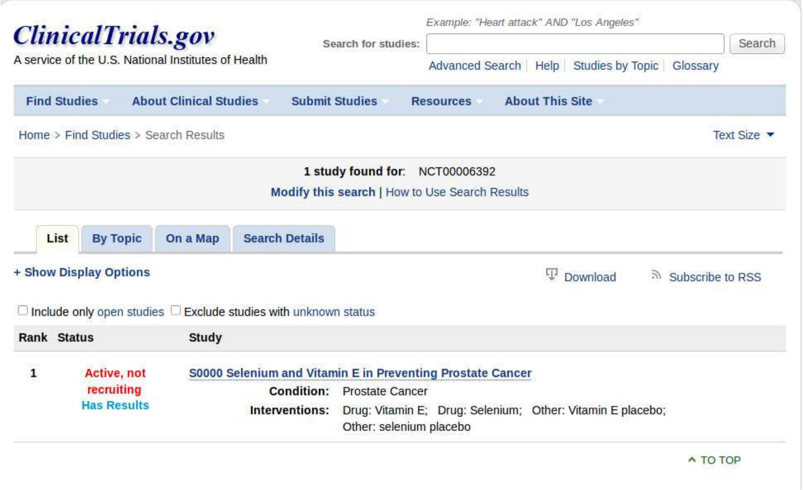
As a result of marking, every subtype of medical intervention aggregated the following number of the abstracts: Behavioral – 585, Biological – 242, Device – 238, Dietary Supplement – 191, Drug – 1619, Other – 333, Procedure – 300, Radiation – 18, and Genetic – 8. After that, 3534 abstracts were divided into training set and testing set for performing training and testing of the following classifiers: Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM using the sklearn library. Also, experiments with changing parameters of classifiers were performed in order to determine the most efficient algorithm for classification.

IV. Experimental part

This section presents experiment results for multinomial classification of automatically marked 3534 MEDLINE abstracts by subtypes of medical interventions. Such algorithms as Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM were used (names of algorithms in sklearn library: MultinomialNB, LinearSVC, LogisticRegression). Marked abstracts were divided into training set and testing set in ascending order by date of publication as follows:

- For training:
  - 2651 abstracts from 2006 till 2012 year containing the following number of classes with abstracts: Behavioral - 450, Biological - 174, Device- 189, Dietary Supplement - 145, Drug - 1198, Other - 266, Procedure - 209, Radiation - 12, Genetic - 8.
- For testing:
  - 883 abstracts from 2012 till 2013 year containing the following number of classes with abstracts: Behavioral - 135, Biological - 68, Device - 49, Dietary Supplement - 46, Drug - 421 Other - 67, Procedure - 91, Radiation - 6.

As a partition result, testing set contained the following sub-types of medical interventions: Behavioral, Biological,



1 study found for: NCT00006392

Modify this search | How to Use Search Results

□ Include only open studies □ Exclude studies with unknown status

Rank	Status	Study
1	Active, not recruiting Has Results	S0000 Selenium and Vitamin E in Preventing Prostate Cancer Condition: Prostate Cancer Interventions: Drug: Vitamin E; Drug: Selenium; Other: Vitamin E placebo; Other: selenium placebo

Fig. 1. Search results by link at Cliniclatrials.gov

Dietary Supplement, Drug, Other, Procedure, and Radiation. To evaluate classification result for the testing set for each subtype of medical intervention such measures as precision, recall and F-measure were used:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall},$$

where TP – true positive classification value, i.e. the classifier identified an element of the testing set correctly. FP – false positive value, i.e. the classifier referred an element to the class falsely. FN – false negative elements, i.e. the classifier falsely did not refer an element to the class. To assess different multinomial classification algorithms, macro and micro precision and recall values, as well as F-measure were calculated using the following formulas:

Macro:

$$precision = \frac{\sum_{n=1}^{m} Precision_n}{m},$$

$$recall = \frac{\sum_{n=1}^{m} Recall_n}{m},$$

$$F\text{-}measure = 2 * \frac{precision * recall}{precision + recall}.$$

Micro:

$$precision = \frac{\sum_{n=1}^{m} TP_n}{\sum_{n=1}^{m} TP_n + \sum_{n=1}^{m} FP_n},$$

$$recall = \frac{\sum_{n=1}^{m} TP_n}{\sum_{n=1}^{m} TP_n + \sum_{n=1}^{m} FN_n},$$

$$F\text{-}measure = 2 * \frac{(precision) * (recall)}{(precision) + (recall)},$$

where m is the number of classes. Calculation of micro precision and recall was performed by summing up all true positive, false positive and false negative results of classification for each class.

In this experiment we use the «bag of words» model [9]. For vector of features we use vector of terms from the dictionary, composed of all annotations from corpora. In process of forming dictionary no stemming was used. With the help of $tf\text{-}idf$ metric, weight for every term was assessed.

$$tf\text{-}idf(t, d) = tf(t, d) * idf(t),$$

where tf – term frequency:

$$tf(t, d) = \frac{n_{t,d}}{|d|},$$

t – term, d – document, $n_{t,d}$ – entry of t term occurence in d document, $|d|$ – total number of terms in d document; idf – inverse document frequency:

$$idf(t) = \log \frac{N}{df(t)},$$

where N – number of documents in corpora; df – number of documents, in which term t occures.

During the experiments, the optimal parameters of classifiers were selected for the case with the removal of stop words, as well as for the case with no stop words removing from the dictionary. More detailed options of the algorithms and their values are given in the documentation for the library

TABLE IV
RESULT OF CLASSIFICATION WITH THE REMOVAL OF STOP WORDS

	Precision			**Recall**			**F-measure**		
	Multinomial Naive Bayes	**Linear SVM**	**Maximum Entropy**	**Multinomial Naive Bayes**	**Linear SVM**	**Maximum Entropy**	**Multinomial Naive Bayes**	**Linear SVM**	**Maximum Entropy**
Behavioral	77%	73%	70%	49%	78%	76%	60%	75%	73%
Biological	100%	89%	96%	24%	72%	65%	38%	80%	77%
Device	0%	62%	71%	0%	41%	10%	0%	49%	18%
Dietary Supplement	0%	58%	58%	0%	63%	24%	0%	60%	34%
Drug	53%	79%	65%	99%	94%	97%	69%	86%	78%
Other	100%	34%	50%	1%	22%	9%	3%	27%	15%
Procedure	0%	64%	65%	0%	42%	14%	0%	51%	23%
Radiation	0%	100%	0%	0%	33%	0%	0%	50%	0%

TABLE V
RESULT OF CLASSIFICATION WITH THE REMOVAL OF STOP WORDS (MICRO AND MACRO EVALUATION)

	Precision		**Recall**		**F-measure**	
	Micro	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**
Multinomial Naive Bayes	55.77%	41.27%	86.87%	21.62%	67.93%	28.37%
Linear SVM	64.91%	70.06%	79.13%	55.62%	71.32%	62.01%
Maximum Entropy	65.66%	59.34%	76.63%	36.93%	70.72%	45.53%

TABLE VI
RESULT OF CLASSIFICATION WITHOUT STOP WORDS REMOVAL

	Precision			**Recall**			**F-measure**		
	Multinomial Naive Bayes	**Linear SVM**	**Maximum Entropy**	**Multinomial Naive Bayes**	**Linear SVM**	**Maximum Entropy**	**Multinomial Naive Bayes**	**Linear SVM**	**Maximum Entropy**
Behavioral	87%	73%	74%	1%	84%	76%	17%	78%	75%
Biological	0%	89%	96%	0%	75%	65%	6%	82%	77%
Device	0%	56%	67%	0%	39%	8%	0%	46%	15%
Dietary Supplement	0%	59%	57%	0%	70%	26%	0%	64%	36%
Drug	48%	80%	65%	100%	93%	97%	65%	86%	78%
Other	0%	59%	63%	0%	38%	13%	0%	47%	22%
Procedure	0%	100%	0%	0%	33%	0%	0%	50%	0%
Radiation	0%	100%	0%	0%	33%	0%	0%	50%	0%

TABLE VII
RESULT OF CLASSIFICATION WITHOUT STOP WORDS REMOVAL (MICRO AND MACRO EVALUATION)

	Precision		**Recall**		**F-measure**	
	Micro	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**
Multinomial Naive Bayes	49.03%	16.88%	77.87%	13.67%	60.18%	15.10%
Linear SVM	61.72%	69.72%	75.74%	56.73%	68.01%	62.56%
Maximum Entropy	63.49%	57.87%	72.62%	37.19%	37.19%	45.28%

sklearn. Below the optimum parameters of the algorithms are presented.

The following algorithm parameters were used for the classification of documents from the test set with no account for stop words:

MultinomialNB: alpha = 1.0, fit_prior = True;

LinearSVC: penalty = '12', loss = 'squared_hingle', multi_class = 'ovr', C = 1.0;

LogisticRegression: penalty = '12', multi_class = 'ovr', C = 1.0, solver = 'liblinear';

Results are presented in tables IV, V.

The following algorithm parameters were used for the classification of the testing set with stop words in the dictionary:

MultinomialNB: alpha = 2.0, fit_prior = True;

LinearSVC: penalty = '11', loss = 'hingle', multi_class = 'crammer_singer', C = 0.8;

LogisticRegression: penalty = '11', multi_class = 'multinomial', C = 0.8, solver = 'newton-cg'.

Corresponding results are presented in tables VI, VII.

Based on the results of computational experiments, the best results were obtained without accounting for stop words in the dictionary and when using LinearSVC with the following parameters: penalty = '12', loss = 'squared_hingle', multi_class = 'ovr', C = 1.0.

Corresponding results for macro evaluation: precision= 70.06%; recall = 55.62%; F-measure = 62.01%. Results of micro evaluation: precision = 64.91%; recall= 79.13%; F-measure = 71.32%.

V. Discussion

Relatively low classification quality rates are associated with the fact, that documents for classification describe medical studies, which were performed during patients treatment. Some differences in documents from various classes are related only to subtypes of medical treatments, that were considered in the studies, and can describe patients suffering from the same disease. The result that was recieved can be compared with results from the research [10],where maximum F-measure value of 80% has been achieved by using linear SVM during the classification of abstracts on RCTs and on non RCTs. In our case maximum value of macro F-measure = 62.01% and micro F-measure = 71.32% has been also retrieved when using linear SVM, with multiclassification of abstracts by subtypes of medical interventions.

VI. Conclusion

This article describes methods that allow to automate grouping MEDLINE abstracts by subtypes of medical interventions. Computational experiments were carried out using the following algorithms: Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM from the sklearn library. Linear SVM algorithm showed the best result of multinomial classification.

For further research it is planned to perform the following tasks:

- chose the set of nonlinear multinomial classifier algorithms and examine 3534 MEDLINE abstracts using these algorithms;
- classify the remaining 1996466 unmarked abstracts using Linear SVM algorithm;
- extract facts from the marked abstracts about a specific subtype of a medical intervention described in the study;
- group abstracts by subtypes of medical intervention using a catalog of natural science subjects MESH[5] and contents of the <topics> field from the document structure (1).

References

[1] G. Guyatt,G. Vist, Y. Falck-Ytter, R. Kunz, N. Magrini ,H. Schunemann for the GRADE* working group. "An emerging consensus on grading recommendations?," (Editorial). *ACP J Club,* 2006, Jan-Feb;144(1):A08, PMID: 17216711.

[2] G. Guyatt,G. Vist, Y. Falck-Ytter, R. Kunz, N. Magrini ,H. Schunemann for the GRADE* working group."GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables,"*Journal of Clinical Epidemiology,* 2011, vol. 64, pp. 383-394, doi: 10.1016/j.jclinepi.2010.04.026.

[3] T. Ohta, Y. Tsuruoka, J. Takeuchi, J. Kim, Y. Miyao, A. Yakushiji et al. "An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing," *Proceedings of the COLING/ACL on Interactive presentation sessions,* Stroudsburg, PA, USA, 2006, vol. 4, pp. 17-20, doi: 10.3115/1225403.1225408.

[4] S. Kaneko, A. Hayashi, N. Suematsu, K. Iwata, "Hierarchical hidden conditional random fields for information extraction," *Proceedings of the 5th international conference on Learning and Intelligent Optimization,* Springer-Verlag Berlin, Heidelberg, 2011, vol. 12, pp. 191-202, doi: 10.1007/978-3-642-25566-3_14.

[5] A. Hliaoutakis, K. Zervanou, E. G.M. Petrakis, E. E. Milios, "Automatic document indexing in large medical collections," *Proceedings of the international workshop on Healthcare information and knowledge management*, New York, USA, 2006, vol. 8, pp. 1-8, doi: 10.1145/1183568.1183570.

[6] V. Dobrynin, D. Patterson, M. Galushka, N. Rooney, "SOPHIA: An Interactive Cluster Based Retrieval System for the OHSUMED collection," *in IEEE Trans. on Information Technology for Biomedicine*, 2005 , vol. 9, pp. 256-265, PMID: 16138542 .

[7] K. Veningston, R. Shanmugalakshmi, "Information Retrieval by Document Re-ranking using Term Association Graph," *Proceedings of the 2014 International Conference on Interdisciplinary Advances in Applied Computing*, New York, USA, 2014 , vol. 8, Article No. 21., doi:10.1145/2660859.2660927

[8] D. Demner-Fushman, J. Lin, "Answering Clinical Questions with Knowledge-Based and Statistical Techniques," *Journal of Computational Linguistics*, 2007 , vol. 33, pp. 63-103, doi: 10.1162/coli.2007.33.1.63

[9] Ch. D. Manning, P. Raghavan, H. Schutze, "Introduction to Information Retrieval," *Cambridge University Press*, Cambridge, England, 2008 , pp. 482, isbn: 9780521865715

[10] A. M. Cohen , N. R. Smalheiser , M. S. McDonagh , C. Yu , C. E. Adams , J. M. Davis et al. "Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine," *Journal of Am Med Inform Assoc* , 2015, pp. 707-717, doi: http://dx.doi.org/10.1093/jamia/ocu025

[11] E. Afful-Dadzie , S. Nabareseh , S. Kominkova Oplatkova, "Fuzzy VIKOR approach: evaluating quality of internet health information," *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems* , Warsaw, Poland, 2014, vol. 2, pp. 183-190, doi: 10.15439/2014F203

[5] http://www.nlm.nih.gov/mesh/