International Journal of Risk & Safety in Medicine 27 (2015) S106–S107
DOI 10.3233/JRS150710
IOS Press

Improving data retrieval quality: Evidence based medicine perspective

M. Kamalov[a,*], V. Dobrynin[a], J. Balykina[b], A. Kolbin[b], E. Verbitskaya[c] and M. Kasimova[c]

[a]*Saint-Petersburg State University, Saint-Petersburg, Russia*
[b]*Pavlov First Saint-Petersburg State Medical University, Saint-Petersburg, Russia*
[c]*Tashkent Institute of Postgraduate Medical Education, Tashkent, Uzbekistan*

*Corresponding author. E-mail: mkamalovv@gmail.com

BACKGROUND: The actively developing approach in modern medicine is the approach focused on principles of evidence-based medicine. The assessment of quality and reliability of studies is needed. However, in some cases studies corresponding to the first level of evidence may contain errors in randomized control trials (RCTs). Solution of the problem is the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system. Studies both in the fields of medicine and information retrieval are conducted for developing search engines for the MEDLINE database [1]; combined techniques for summarization and information retrieval targeted to solving problems of finding the best medication based on the levels of evidence are being developed [2].

OBJECTIVE: Based on the relevance and demand for studies both in the field of medicine and information retrieval, it was decided to start the development of a search engine for the MEDLINE database search 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. Novelty and value of the proposed system are characterized by the use of ranking method of relevant abstracts. It is suggested that the system will be able to perform ranking based on studies level of evidence and to apply GRADE criteria for system evaluation.

METHODS: The assigned task falls within the domain of information retrieval and machine learning. Based on the results of implementation from previous work [3], in which the main goal was to cluster abstracts from MEDLINE database by subtypes of medical interventions, a set of algorithms for clustering in this study was selected: K-means, K-means ++, EM from the sklearn (http://scikit-learn.org**)** and WEKA (http://www.cs.waikato.ac.nz/~ml/weka/) libraries, together with the methods of Latent Semantic Analysis (LSA) [4] choosing the first 210 facts and the model "bag of words" [5] to represent clustered documents. During the process of abstracts classification, few algorithms were tested including: Complement Naive Bayes [6], Sequential Minimal Optimization (SMO) [7] and non linear SVM from the WEKA library.

RESULTS: The first step of this study was to markup abstracts of articles from the MEDLINE by containing and not containing a medical intervention. For this purpose, based on our previous work [8] a web-crawler was modified to perform the necessary markuping. The next step was to evaluate the clustering algorithms at the markup abstracts. As a result of clustering abstracts by two groups, when applying the LSA and choosing first 210 facts, the following results were obtained:

1) K-means: Purity = 0,5598, Normalized Entropy = 0.5994;
2) K-means ++: Purity = 0,6743, Normalized Entropy = 0.4996;
3) EM: Purity = 0,5443, Normalized Entropy = 0.6344.

When applying the model "bag of words":

1) K-means: Purity = 0,5134, Normalized Entropy = 0.6254;
2) K-means ++: Purity = 0,5645, Normalized Entropy = 0.5299;
3) EM: Purity = 0,5247, Normalized Entropy = 0.6345.

Then, studies which contain medical intervention have been considered and classified by the subtypes of medical interventions. At the process of classification abstracts by subtypes of medical interventions, abstracts were presented as a "bag of words" model with the removal of stop words. The results:

1) Complement Naive Bayes: macro F-measure = 0.6934, micro F-measure = 0.7234;
2) Sequantial Minimal Optimization: macro F-measure = 0.6543, micro F-measure = 0.7042;
3) Non linear SVM: macro F-measure = 0.6835, micro F-measure = 0.7642.

CONCLUSIONS: Based on the results of computational experiments, the best results of abstract clustering by containing and not containing medical intervention were obtained using the K-Means ++ algorithm together with LSA, choosing the first 210 facts. The quality of classification abstracts by subtypes of medical interventions value for existing ones [8] has been improved using non linear SVM algorithm, with "bag of words" model and the removal of stop words. The results of clustering obtained in this study will help in grouping abstracts by levels of evidence, using the classification by subtypes of medical interventions and it will be possible to extract information from the abstracts on specific types of interventions.

Keywords: Data retrieval, computational experiments

References

[1] T. Ohta, Y. Tsuruoka, J. Takeuchi, J. Kim, Y. Miyao, A. Yakushiji, K. Yoshida, Y. Tateisi, T. Ninomiya, K. Masuda, T. Hara, J. Tsujii, "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

[2] D. Demner-Fushman, J. Lin, "Answer extraction, semantic clustering, and extractive summarization for clinical question answering," *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Stroudsburg, USA, 2006, pp. 841-848, doi: 10.3115/1220175.1220281

[3] V. Dobrynin, J. Balykina, M. Kamalov, "**Analysis of standard clustering algorithms for grouping MEDLINE abstracts into evidence-based medicine intervention categories**," (**Editorial**) *Proceedings of the III International Conference in memory of V.I.Zubov Stability and Control Processes*, Saint-Petersburg, Russia, 2015.

[4] T. Kolda, D. O'Leary, "A semidiscrete matrix decomposition for latent semantic indexing information retrieval," *Journal of ACM Transactions on Information Systems*, New York, USA, 1998, vol. 16, pp. 322-346, doi: 10.1145/291128.291131

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

[6] J. D. M. Rennie , L. Shih , J. Teevan , D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," *Proceedings of the Twentieth International Conference on Machine Learning,* 2003, pp. 616-623

[7] S. S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design," *Journal Neural Computation,* USA, 2001, vol. 13, pp. 637–649, doi: 10.1162/089976601300014493

[8] V. Dobrynin, J. Balykina, M. Kamalov, A. Kolbin, E. Verbitskaya, M. Kasimova "The data retrieval optimization from the perspective of evidence-based medicine," (**Editorial**) *Proceedings of the Federated Conference on Computer Science and Information Systems 2015*, Lodz, Poland, 2015.