Quality Assessment of Wikipedia Articles: A Deep Learning Approach

Quang-Vinh Dang, Claudia-Lavinia Ignat

Quality Assessment of Wikipedia Articles: A Deep Learning Approach

Quang Vinh Dang and Claudia-Lavinia Ignat
LORIA, Université de Lorraine / Inria / CNRS

Wikipedia is indeed a very important knowledge sharing platform. However, since its start in 2001, the quality of Wikipedia is questioned because its content is created potentially by everyone who can access to the Internet. Currently, the quality of Wikipedia articles is assessed by human judgement. The method is not scalable up to huge size and fast changing speed of Wikipedia today. An automatic quality classifier for Wikipedia articles is required to support users to choose high quality articles for reading and to notify authors for improving their products. While other existing approaches are based on manually predefined specific feature set, we present our approach of using deep learning to automatically represent Wikipedia articles for quality classification.

1. INTRODUCTION

Today, Wikipedia is the largest[1] and probably the most important knowledge repository in the world. At the time of writing, there are more than five millions articles in English Wikipedia particularly, and around 40 millions article in Wikipedia all languages[2]. On average, ten edits per second are performed on Wikipedia[3].

The question related to Wikipedia quality was raised since its start in 2001 [Denning et al. 2005]. Some studies [Holman Rector 2008] claimed that the quality of Wikipedia is not comparable to other traditional encyclopedias.

In order to improve the overall quality of Wikipedia, Wikipedia articles need to be classified based on their quality such that readers can be guided to high quality writing while authors can be notified about low quality texts that require improvements. Several quality classes were defined, from *FA* as the highest quality class to *Stub* as the lowest quality class as shown in Table I. Currently, the quality labels are assigned to articles by human judgement[4]. However, due to the huge size of Wikipedia and fast speed of edits, humans cannot manually review them. Therefore, an automatic quality classifier is required.

Existing approaches rely on feature engineering requiring a manual specification of a fea-

[1] https://en.wikipedia.org/wiki/Wikipedia:Largest_encyclopedia
[2] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
[3] https://tools.wmflabs.org/wmcounter/
[4] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Years/Assessment
[5] https://en.wikipedia.org/wiki/Template:Grading_scheme

Class	Description
FA	Professional, outstanding, and thorough; a definitive source for encyclopedic information.
GA	Useful to nearly all readers, with no obvious problems; approaching (but not equalling) the quality of a professional encyclopedia.
B	Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.
C	Useful to a casual reader, but would not provide a complete picture for even a moderately detailed study.
Start	Provides some meaningful content, but most readers will need more.
Stub	Provides very little meaningful content; may be little more than a dictionary definition. Readers probably see insufficiently developed features of the topic and may not see how the features of the topic are significant.

Table I. Description of quality classes of Wikipedia articles[5].

ture set. However, there is no guarantee that the feature set is complete and usually selection of features depends on researchers' expertise.

2. RELATED WORKS

Existing approaches rely on specific feature set to determine quality of Wikipedia articles. In other words, each approach proposed a set of features that are believed for correlation with articles quality.

One of the first efforts belongs to the work of [Blumenstock 2008] where authors used length of articles as an indicator for their quality. [Dalip et al. 2009] analyzed the effect of the feature set comprising text, review and network on the quality of Wikipedia articles. [Anderka et al. 2012] built a classifier to detect vandalism on Wikipedia based on *cleanup tags*. [Warncke-Wang et al. 2013] analyzed a set of 17 features and eventually came up with 11 features and applied random forest technique for the classification. Results are discussed in more details in [Warncke-Wang et al. 2015].

Based on the work of [Warncke-Wang et al. 2015; Warncke-Wang et al. 2013], Wikimedia Foundation[6] built an online API to predict the quality class of Wikipedia articles called ORES(*Objective Revision Evaluation Service*) [Halfaker and Taraborelli 2015]. The service uses an extended version of the feature set presented by Warncke-Wang et al., with an adaption for French Wikipedia. Finally, Wikimedia Foundation development team used a feature set which includes 24 features for English Wikipedia, and 25 features for French Wikipedia. To our knowledge, Wikimedia ORES service is the state-of-the-art approach in classifying the quality of Wikipedia articles.

Several studies defined features based on information about authors rather than about articles themselves. [Adler et al. 2008] used authors *reputation* as feature set. [Suzuki 2015] applied the concept of *h-index* in academic publication for determining quality of

[6]`https://wikimediafoundation.org`

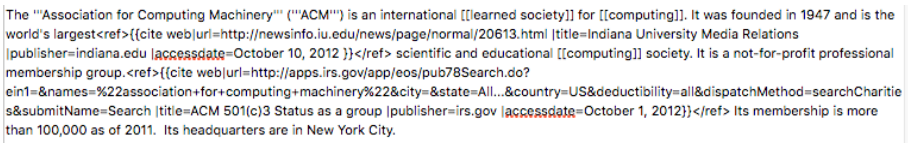

Fig. 1. An example of raw Wikipedia content

Wikipedia articles. Article quality can be also determined by the interaction between authors and reviewers [Hu et al. 2007; Wu et al. 2012; de La Robertie et al. 2015] or based on the network structure of authors and articles [Li et al. 2015].

However, the definition of the feature set used by existing approaches for automatic classification of Wikipedia articles quality is based on the expertise, experience and knowledge of each research team. There is no *gold* standard theory to find the best feature set. When a new feature is defined, the only validation method is empirical testing. For instance, [Halfaker and Taraborelli 2015] used the division of number of images by article length as a feature. But there is no evidence that adding some derived features such as division of number of images by square of content length, or by square root of content length would change the performance of the classifier.

3. A DEEP-LEARNING BASED APPROACH

While traditional machine learning techniques such as *k-NN* or *random forest* [Kubat 2015] require manual feature engineering, deep learning techniques can be used to learn features automatically from the dataset rather than defining them before-hand [LeCun et al. 2015]. We proposed an approach that uses *Doc2Vec* [Le and Mikolov 2014] for learning features from textual documents and Deep Neural Networks [Goodfellow et al. 2016] for classifying Wikipedia articles [Dang and Ignat 2016].

3.1 Method

We used Doc2Vec [Le and Mikolov 2014] and Deep Neural Networks [Goodfellow et al. 2016]. Our proposed approach includes two steps. In the first step we applied Doc2Vec on Wikipedia articles. The idea of Doc2Vec is to convert a variable-length textual document into a fixed-length numerical vector. Instead of applying Doc2Vec on textual documents as the original method presented in [Le and Mikolov 2014], we applied Doc2Vec on *raw* content of Wikipedia articles. This raw content contains all the necessary information to determine the quality of an article. An example of a raw Wikipedia content is provided in Fig. 1.

In the second step we applied Deep Neural Networks (DNN) on output vectors of Doc2Vec. DNN is defined as an artificial neural networks [McCulloch and Pitts 1943] with multiple hidden layers [Goodfellow et al. 2016], as visualized in Fig. 2[7]. DNN receives input values from its input layer, passes the computation through hidden layers and generates predicting values at output layer as follows: the k^{th} layer computes an output vector h^{k}

[7]The image is from [Nielsen 2015] with the permission of usage.

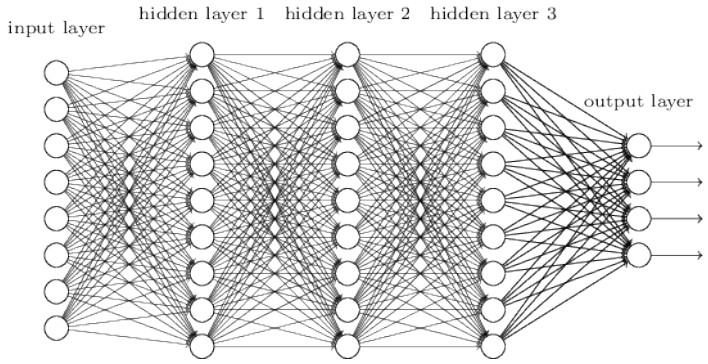

Fig. 2. An example of deep neural networks with three hidden layers

using the output h^{k-1} of the previous layer, starting with the input layer $x = h^0$ as in Equation 1, where b^k is the offset vector and W^k is the matrix of weights.

$$h^k = f(b^k + W^k h^{k-1}) \tag{1}$$

The function f in Equation 1 is called *activation function*, which decides how each neuron calculates and transfers the signal to the neurons in subsequent layer. We used *rectifier* as our activation function, defined as follows:

$$rectifier(x) = max(0, x) \tag{2}$$

The values of b and W are set during training process, using *gradient-descent* technique [Bengio 2012].

DNN is an emerging research field, and there are still a lot of open theoretical questions, especially in choosing hyper-parameters for the network [Goodfellow et al. 2016]. In our experiments, we came up with a neural network with four hidden layers as the result of random search technique [Bergstra and Bengio 2012].

3.2 Results

We validated our approach on the public dataset including 30 000 Wikipedia articles provided by Wikimedia Foundation Research[8]. Following the train/test division of existing studies [Warncke-Wang et al. 2015; Halfaker and Taraborelli 2015], we used 80% of the dataset for training and 20% for testing.

For the classification of Wikipedia articles according to all six quality classes, our method achieved the accuracy of 55% compared to the accuracy of 58% of [Warncke-Wang et al. 2015] and 60% of [Halfaker and Taraborelli 2015]. We believe further optimisations can be achieved using the DNN technique.

[8] http://datasets.wikimedia.org/public-datasets/enwiki/

For the binary classification, our method achieved very good results. The method classifies *FA vs Start* with the accuracy of 99%, *FA-GA vs all* with the accuracy of 86% and *FA-GA vs C-Start* of 90%, much higher than existing approaches. For instance, [Xu and Luo 2011] classifies *FA vs Start* with the accuracy of 84%, [Lex et al. 2012] classifies *FA-GA vs all* with the accuracy of 84%, and [Wu et al. 2012] classifies *FA-GA vs C-Start* with the accuracy of 66%.

4. CONCLUSIONS

Feature engineering is a norm in machine learning for a long time [Kubat 2015] with a lot of hands-on tips which are accumulated through a lot of works [Zheng 2016]. The recent research on deep learning brings a new potential to machine learning field to build a system which can learn to represent the data by itself. Our work presented a new way to understand the problem of Wikipedia articles classification, which is an important topic in both research and practice.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Aaron Halfaker, Wikimedia Research Foundation and Morten-Warncke Wang, University of Minnesota for providing the dataset and valuable discussions.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

REFERENCES

ADLER, B. T., CHATTERJEE, K., DE ALFARO, L., FAELLA, M., PYE, I., AND RAMAN, V. 2008. Assigning trust to wikipedia content. In *Int. Sym. Wikis*. ACM.

ANDERKA, M., STEIN, B., AND LIPKA, N. 2012. Predicting quality flaws in user-generated content: the case of wikipedia. In *SIGIR*. ACM, 981–990.

BENGIO, Y. 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade (2nd ed.)*. Lecture Notes in Computer Science, vol. 7700. Springer, 437–478.

BERGSTRA, J. AND BENGIO, Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13*, 281–305.

BLUMENSTOCK, J. E. 2008. Size matters: word count as a measure of quality on wikipedia. In *WWW*. ACM, 1095–1096.

DALIP, D. H., GONÇALVES, M. A., CRISTO, M., AND CALADO, P. 2009. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *JCDL*. ACM, 295–304.

DANG, Q. V. AND IGNAT, C. 2016. Quality assessment of wikipedia articles without feature engineering. In *JCDL*. ACM, 27–30.

DE LA ROBERTIE, B., PITARCH, Y., AND TESTE, O. 2015. Measuring article quality in wikipedia using the collaboration network. In *ASONAM*. ACM, 464–471.

DENNING, P. J., HORNING, J., PARNAS, D. L., AND WEINSTEIN, L. 2005. Wikipedia risks. *Commun. ACM 48,* 12, 152.

GOODFELLOW, I., BENGIO, Y., AND COURVILLE, A. 2016. *Deep Learning*. MIT Press.

HALFAKER, A. AND TARABORELLI, D. 2015. Artificial intelligence service gives Wikipedians 'x-ray specs' to see through bad edits. https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs. Accessed: 2016-04-01.

HOLMAN RECTOR, L. 2008. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review 36,* 1, 7–22.

HU, M., LIM, E., SUN, A., LAUW, H. W., AND VUONG, B. 2007. Measuring article quality in wikipedia: models and evaluation. In *CIKM*. ACM, 243–252.

KUBAT, M. 2015. *An Introduction to Machine Learning*. Springer.

LE, Q. V. AND MIKOLOV, T. 2014. Distributed representations of sentences and documents. In *ICML*. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 1188–1196.

LECUN, Y., BENGIO, Y., AND HINTON, G. 2015. Deep learning. *Nature 521,* 7553, 436–444.

LEX, E., VOELSKE, M., ERRECALDE, M., FERRETTI, E., CAGNINA, L., HORN, C., STEIN, B., AND GRANITZER, M. 2012. Measuring the quality of web content using factual information. In *Proc. of WICOW*. 7–10.

LI, X., TANG, J., WANG, T., LUO, Z., AND DE RIJKE, M. 2015. Automatically assessing wikipedia article quality by exploiting article-editor networks. In *ECIR*. Lecture Notes in Computer Science, vol. 9022. 574–580.

MCCULLOCH, W. S. AND PITTS, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics 5,* 4, 115–133.

NIELSEN, M. A. 2015. *Neural networks and deep learning*. http://neuralnetworksanddeeplearning.com.

SUZUKI, Y. 2015. Quality assessment of Wikipedia articles using h-index. *Journal of Information Processing 23,* 1, 22–30.

WARNCKE-WANG, M., AYUKAEV, V. R., HECHT, B., AND TERVEEN, L. G. 2015. The success and failure of quality improvement projects in peer production communities. In *CSCW*. ACM, 743–756.

WARNCKE-WANG, M., COSLEY, D., AND RIEDL, J. 2013. Tell me more: an actionable quality model for wikipedia. In *OpenSym*. ACM, 8:1–8:10.

WU, G., HARRIGAN, M., AND CUNNINGHAM, P. 2012. Classifying wikipedia articles using network motif counts and ratios. In *WikiSym*. ACM, 12.

XU, Y. AND LUO, T. 2011. Measuring article quality in Wikipedia: Lexical clue model. In *Proc. of SWS*. 141–146.

ZHENG, A. 2016. *Mastering Feature Engineering: Principles and Techniques for Data Scientists*. O'Reilly Media.

Quang-Vinh Dang is a PhD candidate at Université de Lorraine, Nancy, France. He conducts researches in several topics of collaboration networks, including quality assessment, users' behavior and trust relationship using deep learning techniques.

Claudia-Lavinia Ignat is a researcher at Inria in France. She has an expertise in distributed collaborative systems with a focus on consistency maintenance, group awareness, security and trust issues and user studies. She investigates computational trust models based on the quality of user contributions during collaboration.