Content-based recommendations in e-commerce services

John Smith*

Warsaw University of Technology
Faculty of Mathematics and Information Science
ul. Koszykowa 75
00-662 Warsaw, Poland
john@smith.com

February 21, 2018

Abstract

Recommendation systems play an important role in modern e-commerce services. The more relevant items are presented to the user, the more likely s/he is to stay on a website and eventually make a transaction. In this paper, we adapt some state-of-the-art methods for determining similarities between text documents to content-based recommendations problem. The aim is to investigate a possibility of improving the existing recommendation system being a part of Allegro e-commerce platform using semantic text analysis methods. As a conclusion we show, that there is no significant difference between examined methods and previously used elasticsearch-based query in content-based recommendation task.

Keywords: recommendations, natural language processing, word embedding, semantics, allegro

I. Introduction

Recommendation systems are commonly used in e-commerce services. Such a system gives profit to both the user, allowing him to get to the information he is he could not find, or know, that such information exists, as well as the owner website, which depends on attracting users to you they used his services to the greatest extent. The key issue of recommendations generation is how suggested items are relevant to these which the user is interested in.

We can divide recommendation systems into two groups: collaborative and content-based filtering. The first one assumes that user is likely to be interested in items which also users similar to s/he were interested in. In this paper we are focusing on the second group in which recommended items are similar to these that the user liked so far.

*A thank you or further information

A detailed problem comes from the on-line auction site Allegro. Allegro — the biggest marketplace platform in Eastern Europe — contains a section presenting text articles concerning products available via the platform. Currently there is a list of links to articles similar to given one. Using elasticsearch

The subject of this paper focuses on issues of determining a semantic similarity between text documents and recommendations of similar documents. In this paper we strive to check if newly proposed word embeddings methods are able to replace recently used method based on Elasticsearch query.

In the following sections we describe topics in detail...

II. Dataset

Given dataset consists of 20000 text articles describing several categories of products available via Allegro platform. Single record consists of

article contents and metadata attached by an author of the article.

Id, category, keywords.

All articles are written in Polish. The vocabulary based on the articles set contains many industry-specific words like brands and models names, books titles and technical words.

Moreover raw articles contain some tags responsible for images and hyperlinks present on the website.

III. Methods

i. Preprocessing

The following describes the next steps of the pre-processing of text that he performs on his own collection of articles.

1. Cleansing the text from the redundant, previously mentioned markers. From the viewpoint of semantic analysis of the text they are useless or harmful (they cause some "Pollution" of the text). So I remove them using properly constructed ones regular expressions. An example of such a tag is! [2_new.jpg] (http: // (...) '2_new.jpg') placing the picture in the middle of the text (the content of the URL removed the reasons confidentiality).

2. Removal of stopwords - usually short words that do not mean anything the whole article. They are, for example, "in", "from", "because". Removing them reduces the number of words document, thus shortening the processing time. As they say these words often, removing them gives you the opportunity to emphasize the meaning of other words that have an impact on the real meaning of the whole article. The collection of alloy words is derived from [31].

3. Bringing all the words of the document to lowercase letters. It helps to unify the character parts of words with the same meaning, among which one occurs at the beginning of the sentence, and others in the middle.

4. Breaking down the words connected with the thought. Experience at a later stage (tokenization) shows that the tool making it (Morfologik [20]) does not deal with these types of words and leaves them in the unchanged grammatical form (eg "white and red"). So it is necessary the manual execution of the mechanism by which I break such words into characters compatible with the tokenizer. The earlier one was needed to perform the appropriate function analysis of this type of words for the behavior of the two elements depending on them the type, case and occurrence of specific letters in the suffixes of the component words. depended me not to break down words that are my own names or symbols of devices.

5. Tokenization. This is the most important element of the whole process. It's about bringing words with the same meaning, and different grammatical form to the same form. A big hurdle here is the degree of complexity of the Polish language and the number of exceptions that this one has language has. As an example, you can use the word "have", which one of the forms is "has", the next is "have it". The goal of the stage is to bring each of these words to the basic form "have". To carry out this operation, he uses the Morfologik tool [20].

The above steps lead the data to a state in which you can apply semantic techniques text analysis. The dictionary built on the preprocessed body contains 98174 unique words, and 7409145 all words (with repetitions).

the overwhelming majority of the words of the dictionary built on the body are the words appearing rarely.

ii. Model building

Word2Vec authors offer pretrained model but only for English language. For Polish it's more difficult... Lack of presumably important words like brand and model names etc.

Text requiring further explanation[1].

iii. Evaluation

In order to compare the methods used to determine the similarity between the articles necessary there is the formalization of certain measures of this similarity.

In practice, however, it is rarely worth the value of a given element of the ranking is adequate to the query generating this ranking.

Evaluation of recommendations generation methods is a nontrivial task. A degree of similarity of two articles can be perceived differently by different persons. Therefore we decided to evaluate methods' performance with .... First with each tested method we choose 6 most similar articles for 50 randomly chosen base articles. Next we split them into pairs: base article — similar article; it gave us $5 * 50 = 250$ pairs for a method. After that 5 persons individually evaluated a similarity of pairs of articles giving scores in scale 1-10. Finally we took an average score for each pair and it allowed us to calculate an average score for each method. Besides tested methods we evaluated also previous Allegro method and randomly generated pairs of articles just for a comparison purpose.

IV. Results and discussion

Expert evaluation directly shows that none of the tested methods is significantly different from others. None of the tested adaptations of the methods of semantic text analysis does not depart as well from the current method used in Allegro. The difference between the best method (FastText) and the worst (LDA) is around 9better than the random method.

To answer the question posed at the beginning of this work, i.e. whether the semantic method Natural language analysis adapted to determine recommendations are in the state to match the quality of the current method used in Allegro, he performs the test he has to check whether there are grounds to believe

[1]Example footnote

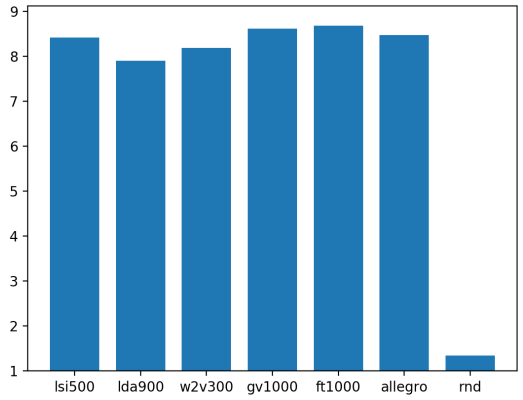

Figure 1

that the results of any of the methods subjected expert judgment is detached statistically significantly from the rest of the tested methods. The statistical test that I carry out is the Kruskal-Wallis test. H0 accepts for the null hypothesis he equality of distributions of distributions in the populations from which the samples originated. In earlier tests, users assessed the similarity between the article and everyone from articles recommended for it, determined by evaluated methods. For data Kruskal-Wallis test entry accepts tests corresponding to test results for each from methods (except for the random method) submitted from the average assessment of users for the most relevant recommendations for each of the tested base articles. In the test, the level of significance is assumed? = 0.05. Finally, as a result of the Kruskal test Wallis receives p = 0.0571>? On the basis of which I state that there are no reasons to reject the null hypothesis - the difference between the quality of the recommendations of the tested methods is not 56 5.5. Test summary statistically significant. It means that neither of the tested semantic methods does not stick out from others, neither the previous Allegro method is significantly better / worse than others tested methods.

V. Conclusion

Comparison of the best configurations of the tested methods showed that there are no signif-

icant differences between methods of semantic text analysis adapted to the task of generating recommendations and the previous method used in Allegro based on the query for the Elasticsearch engine.

After analyzing the test results you can answer: yes, each of the methods of semantic analysis of the natural language in yours the best configuration can be adapted to generate article recommendations based on the content of the article currently displayed by the user of the website. recommendations these in the opinion of users do not undermine the quality of the current method. Using the methods of semantic text analysis allows to capture the hidden similarities between documents, where the documents combine not the same words or synonyms, but some abstract ones concepts related to each other. An important advantage of semantic methods compared to the previous one the method used in Allegro is the fact that they are based only on content articles. This frees the authors of the articles from independently providing them with additional metatags, of which the current method is used to a large extent.

References

[Figueredo and Wolf, 2009] Figueredo, A. J. and Wolf, P. S. A. (2009). Assortative pairing and life history strategy - a cross-cultural study. *Human Nature*, 20:317–330.

[1] Opis Allegro `https://magazyn.allegro.pl/3333-serwis-allegro-to-nasz-sposob-na-wasze-szybkie-i-wygodne-zakupy-przez-internet` (dostęp 07.05.2017)

[2] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, *A Neural Probabilistic Language Model*, Journal of Machine Learning Research 3 1137–1155, 2003

[3] D. M. Blei, A. Y. Ng, M. I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, tom 3 num. 4–5, 2003

[4] Blog Aylien `http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove` (dostęp 18.08.2017)

[5] R. Collobert, J. Weston, *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning*, NEC Labs America, 2008

[6] W. B. Croft, D. Metzler, T. Strohman, *Search Engines. Information Retrieval in Practice*, Pearson Education, Inc., 6-9, 2015

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, tom 41, num. 6, 1990

[8] Elasticsearch `https://www.elastic.co/` (dostęp 18.08.2017)

[9] Firmy korzystające z Elasticsearch `https://www.elastic.co/use-cases` (dostęp 10.08.17)

[10] J.R. Firth, *A synopsis of linguistic theory 1930-1955*, Oxford: Philological Society, 1957

[11] G. H. Golub, W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis. 2 (2), 1965

[12] Z. S. Harris, *Distributional Structure*, Word, 10 (2/3): 146–62, 1954

[13] Informacje o Word2vec `https://code.google.com/archive/p/word2vec/` (dostęp 26.05.2017)

[14] K. Jarvelin, J. Kekalainen, *Cumulated gain-based evaluation of IR techniques*, University of Tampere, 2002

[15] A. Joulin, E. Grave, P. Bojanowski T. Mikolov, *Bag of Tricks for Efficient Text Classification*, Facebook AI Research, 2016

[16] P. Kędzia, G. Czachor, M. Piasecki, J. Kocoń *Vector representations of polish words (Word2Vec method)*, Wrocław University of Technology, 2016, https://clarin-pl.eu/dspace/handle/11321/327 (dostęp 26.06.2017)

[17] M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, *From Word Embeddings To Document Distances*, International Conference on Machine Learning (ICML), 2015

[18] Lucene https://lucene.apache.org/ (dostęp 18.08.2017)

[19] T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*, International Conference on Machine Learning (ICML), 2013

[20] Morfologik http://morfologik.blogspot.com/ (dostęp 07.05.2017)

[21] Opis GloVe https://cran.r-project.org/web/packages/text2vec/vignettes/glove.html (dostęp 30.08.2017)

[22] Opis Word2vec http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/ (dostęp 26.05.2017)

[23] O. Pele, M. Werman, *Fast and robust earth mover's distances*, ICCV, 2009

[24] J. Pennington, R. Socher, C. D. Manning, *GloVe: Global Vectors for Word Representation*, Computer Science Department, Stanford University, Stanford, CA 94305, 2014

[25] Porównanie największych polskich serwisów aukcyjnych http://gadzetomania.pl/11824, zakupy-w-sieci-porownanie-najwiekszych-polskich-serwisow-aukcyjnych-2 (dostęp 09.08.17)

[26] F. Ricci, L. Rokach, B. Shapira, *Introduction to Recommender Systems Handbook*, Springer, 1-4, 2011

[27] F. Ricci, L. Rokach, B. Shapira, *Introduction to Recommender Systems Handbook*, Springer, 145-147, 2011

[28] F. Ricci, L. Rokach, B. Shapira, *Introduction to Recommender Systems Handbook*, Springer, 73-75, 2011

[29] Y. Rubner, C. Tomasi, L. J. Guibas, *The Earth Mover's Distance as a Metric for Image Retrieval*, Computer Science Department, Stanford University, 1, 2000

[30] G. Salton and M. McGill, *Introduction to modern information retrieval*, McGraw-Hill, 1983

[31] Słowa stopu w Wikipedii https://pl.wikipedia.org/wiki/Wikipedia:Stopwords (dostęp 15.04.2017)

[32] Słownik Języka Polskiego PWN http://sjp.pwn.pl/sjp/artykul;2441396.html (dostęp 07.05.2017)

[33] Strona Allegro z artykułem https://allegro.pl/artykul/jaka-farba-dla-alergika-55917/ (dostęp 26.06.2017)

[34] Wordnet http://plwordnet.pwr.wroc.pl/wordnet/ (dostęp 28.06.2017)