Content-based Recommender System for Textual Documents Written in Croatian

Ivana Ćavar, Zvonko Kavran, Natalija Jolić
Neven Anđelović, Ivan Cvitić, Marko Gović
Faculty of transport and traffic sciences,
University of Zagreb
Zagreb, Croatia
ivana.cavar@fpz.hr; zvonko.kavran@fpz.hr, natalija.jolic@fpz.hr
neven1504@gmail.com; ivanfpz@gmail.com; marko.govic@gmail.com

Abstract—**The paper describes a content-based recommender system that classifies textual documents written in Croatian. We describe how documents are pre-processed, including procedures of dimensionality reduction, selection of stop-words and creation of document-term matrix. For the text classification, a combination of *v*-fold cross validation and k - nearest neighbours (*k*NN) methods is used. This way, the 'optimal' value of *k* is firstly analyzed, and the results of *v*-fold cross validation are applied for the selection of value *k*. Results are given in the form of classification error analysis.**

Keywords-text mining; recommender system; k-nearest neighbour; content-based classification; document-term matrix.

I. INTRODUCTION

Search tools are one of the most used tools today. In the environment with different kind of sensors, available information, databases and Internet, the problem is not to find data, but to find useful information among available data and all that in the shortest possible time period and respectively with as little effort as possible along the way. This is one of the main reasons why recommender systems have been developed. They have the effect of guiding users in a personalized way to interesting objects in a large space of possible options. Every day examples for this include offering news articles to on-line newspaper readers, based on a prediction of reader interests, or offering customers of an on-line retailer suggestion about what they might like to buy, based on their past history of purchases and/or product searches.

This paper focuses on classification based recommender system developed for graduate students to aid their learning process by suggesting teaching materials based on a prediction of students' interests and past studying history. This problem is highlighted in multidisciplinary studding areas where students came with different levels of knowledge and different study backgrounds. The paper firstly introduces recommender systems and and a short literature review. After this, pre-processing of documents is described; this is followed by the description of text documents classification. The last section gives more details on results and, finally, we draw the conclusions.

II. RECOMMENDER SYSTEMS

Many areas have embraced recommender systems. There are many benefits from their applications; just some examples of that are the Google News' results with 38% more click-through due to recommendation, Netflix's rented 2/3 of movies based on recommendation, and 35% of Amazon's sales are from recommendation [1].

Recommendation systems use a number of different technologies and are basically classified into two broad groups [2]:

- Content-based systems examine properties of the items recommended;
- Collaborative filtering systems recommend items based on similarity measures between users and/or items. The items recommended to a user are those preferred by similar users.

Recommender systems for text documents written in natural language have become the subject of research for the past few decades. In literature [3] [4], examples of automatic classification of documents where web documents or articles are classified by Naïve Bayes algorithm [5] can be found. Other techniques are also applied for classification of textual documents. Description of library documents classification based on *k* - nearest neighbours algorithm can be found in [6]. Text mining-based recommendation systems assist customer decision making in online product customization, where customers describe their interests in textual format, and, based on captured customers' preferences, recommendations are generated [7]. Other examples can be found in [8] [9][10][11][12][13].

When modelling recommendation systems for text documents written in natural language it is important to carry out pre-processing procedures in order to provide good output results. Text documents considered in this recommendation system are multidisciplinary lecture materials written in Croatian.

III. Pre-processing of text documents

Firstly, it was necessary to represent text documents (strings) in a format suitable for text classification. For this purpose, a vector space model is used (Figure 1).

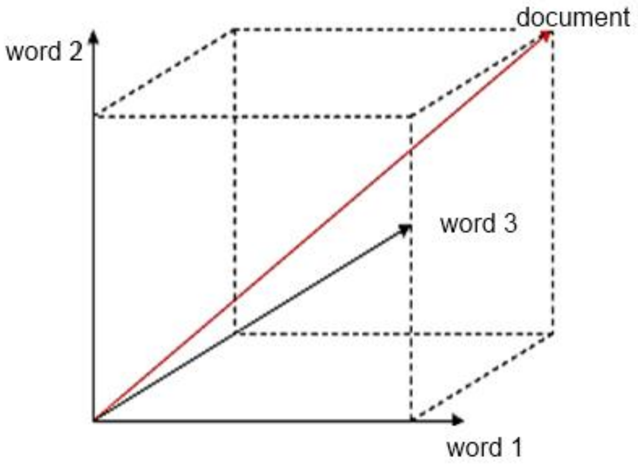

Figure 1. Text documents vector space

In this model, each text document is displayed as a vector of words. So, a document-term matrix or term-document matrix is created (Figure 2).

$$A = \begin{array}{cc} \left[\begin{matrix} a_{11} & \dots & a_{1n} \\ \dots & a_{ij} & \dots \\ a_{m1} & \dots & a_{mn} \end{matrix}\right] & \begin{matrix} w_1 \\ \dots \\ w_i \\ \dots \\ w_m \end{matrix} \\ d_1 \quad \dots \quad d_j \quad \dots \quad d_n & \end{array}$$

Figure 2. Document-term matrix

This is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. For example, $A = (a_{ij})$, where a_{ij} is a weight of word in the document j. There are several ways of determining the weight a_{ij}. Let f_{ij} represent the frequency of words in the document j, N is the number of documents in the learning set, M is the number of different words, and n_i the total number of times a word appears in the learning set. The simplest approach for determining the weight is the binary weights approach, where a_{ij} is set to be *1* if the word appeared in the document; otherwise it is equal to *0*. Another simple method uses the frequency of occurrences of words in a document as a weight, i.e. $a_{ij} = f_{ij}$. The most popular way of determining the weight is Term Frequency - Inverse Document Frequency (TF- IDF) method of determining the weight where the weights are defined as:

$$a_{ij} = f_{ij} \times log\left(\frac{N}{n_i}\right) \qquad (1)$$

If text documents with various lengths are considered, the adopted equation (1) looks like this (since for the matrix A the number of rows is determined by the number of different words in a text document):

$$a_{ij} = \frac{f_{ij}}{\sqrt{\sum_{i=1}^{M} f_{ij}^2}} \times log\left(\frac{N}{n_i}\right) \qquad (2)$$

Given that there may be plenty of different words, including all the words in the language, plus the results of conjugation, and also, gender iterations (as in Croatian different words represent different genders), etc., it was necessary to determine keywords.

For this, the following steps were completed::

- Removing the stop – words;
- Tokenization;
- Lemmatization;
- Stemming;
- Synonyms;
- Group of words;
- Word cleansing.

A. *Stop – words*

Any group of words can be chosen as the stop - words for a given purpose. They can be defined as words that don't have a relevant meaning for the observed subject. Very often, the list of stop- words includes conjunctions, but in some other cases it depends on the document context. The list of stop - words varies depending on the morphological characteristics of the language so that the list for Croatian consists of approximately 2000 words while, for English, this number is approximately 600 words [14].

B. *Tokenization*

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. Tokenization is a very important step in pre-processing documents written in morphologically rich languages like Croatian due to the fact that it allows dimensionality reduction of the input data [15].

C. *Lemmatization*

Lemmatization in linguistics is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is a form of

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-295-0

morphological normalization or procedure that finds the linguistically correct canonical form of a word.

D. *Stemming*

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form, generally, a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words are mapped to the same stem, even if this stem is not in itself a valid root. This type of morphological normalization is less accurate than lemmatization, because the root of the word does not necessarily have to be a meaningful expression [16].

E. *Synonyms*

Synonyms or synonymous words indicate that they have the same meaning. Their identification has a great impact in getting relevant results at the end of text analysis. For example, in Croatian 'ear of corn' has 47 synonyms *(ajdamak, bat, batakljuša, bataljika, batučak, batuček, batuk, baturak, baturice, čepina, čokotinja, ćuka, kic, klas, klasina, klasinec, klasovina, klasovinje, kočanj, kocen, komaljika, komušina, kukuruzina, kumina, kureljica, kuruška, oklipak, oklasak, okoma, okomak, okomina, okrunica, orušek, otučak, paćika, patura, paturica, rucelj, rucl, rulina, šapurika, ščavina, šepurina, štruk, tekun, tulina, tulinek*). Similarly, as for the stop – words, the list is created for the synonyms.

F. *Group of words*

The term group of words represents a problem when a group of individual words, when written together, denote one meaning. For this purpose, we can use two approaches:

- Phrase list - combines word groups that represent common phrases in the language,

- Statistics - monitor the occurrence of two or more words together in the document. Group is defined as group of words if it appeared together more times than a predefined threshold. In order to increase the quality of the text that is the subject of analysis, such a group of words should be represented by one token.

G. *Word cleansing*

Word cleansing process is the last step in the pre-processing procedure. It includes removal of words that appear less frequently. Words that appear less than 1% of the time are usually the result of a typo error and the omission of such words reduces the noise of the document. The same is the case for the words that appear in the document more than 20% of the time.

As an input for document-term matrix creation 54 text documents written in Croatian were used. The used documents are in Portable Document Format (pdf) and selected form teaching materials written in Croatian. The main motivation for selecting these text documents was familiarity with content of this documents and their availability, as well as copy right issues. Based on the results, key words have been extracted for the word set representing each text document. This was done in two phases, manually and automatically. Based on the steps defined in this section, 984 keywords have been identified. These words were used for definition of document-term matrix.

IV. TEXT CLASSIFICATION

For text classification, weighted kNN method is used. When classifying a new document, kNN algorithm needs to determine the closest neighbours by calculating the distance vectors between documents [17]. Based on the *k* most similar neighbour class of considered document is decided. Similarity is determined by the Euclidean distance between vectors of documents or cosine value between two vectors of documents. Cosine value is defined as [18]:

$$sim(X, D_j) = \frac{\sum_{t_j \in (X \cap D_j)} x_i \times d_{ij}}{\left\|X\right\|_2 \times \left\|D_j\right\|_2} \qquad (3)$$

where *X* is the vector of classifying document. D_j is the *j* -th document from the learning set, t_j is a word that exists in *X* and D_j, *x* is the weight of those words in the document D_j, $\|X\| = \sqrt{x_1^2 + x_2^2 + x_3^2 + \ldots}$ is a normal vector *X*, and respectively $\| D_j \|$ is the normal vector D_j.

To determine the optimal size of neighbourhood, *v*-fold cross-validation method was applied. This means that, for a fixed value of *k*, we apply the kNN model to make predictions on the *v*th segment and to evaluate the error. The choice for this error is defined as the accuracy (the percentage of correctly classified cases). This process is then successively applied to all possible choices of *v*. At the end of the *v* folds (cycles), the computed errors are averaged to yield a measure of the stability of the model (how well the model predicts query points). The above steps are then repeated for various *k* and the value achieving the highest classification accuracy is then selected as the value for the *k*. Results of *v*-fold cross-validation in Statsoft Statistica [19] tool are shown in Figure 3.

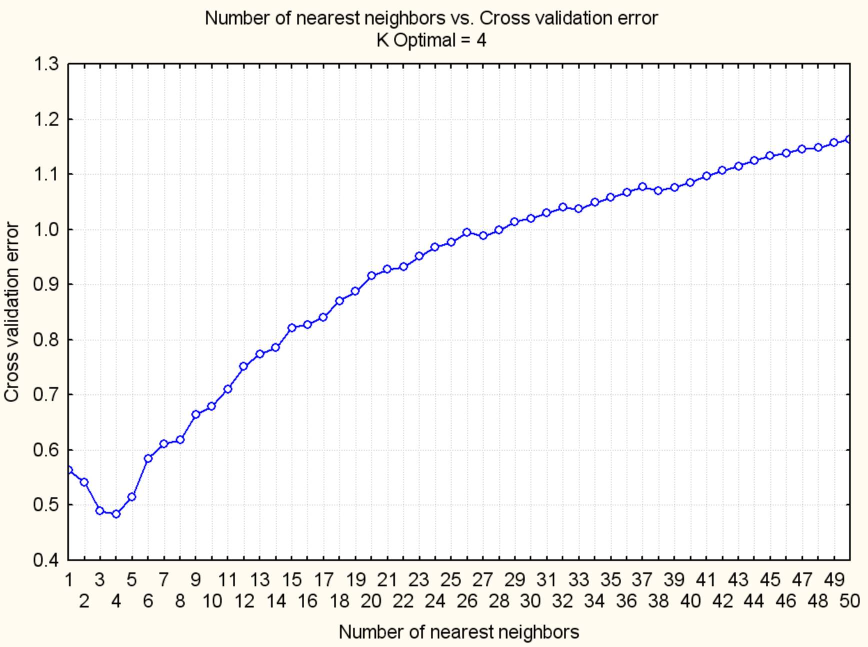

Figure 3. *v*-fold crossvalidation result

Based on the *v*-fold cross-validation results the value of the parameter *k* was set to be 4 (the lowest cross-validation error is 0.48429 %). For the *k* values that are higher than 4, continuous growth of classification errors is recorded. Figure 4 represents classification error where, in multidimensional space, classification error for 10 cases is represented by dots. As visible, most results are in the 'yellow' area, meaning that the value of the error was close to 0. Just one value is in 'red' area, meaning that error was in the interval between 3 - 5 %.

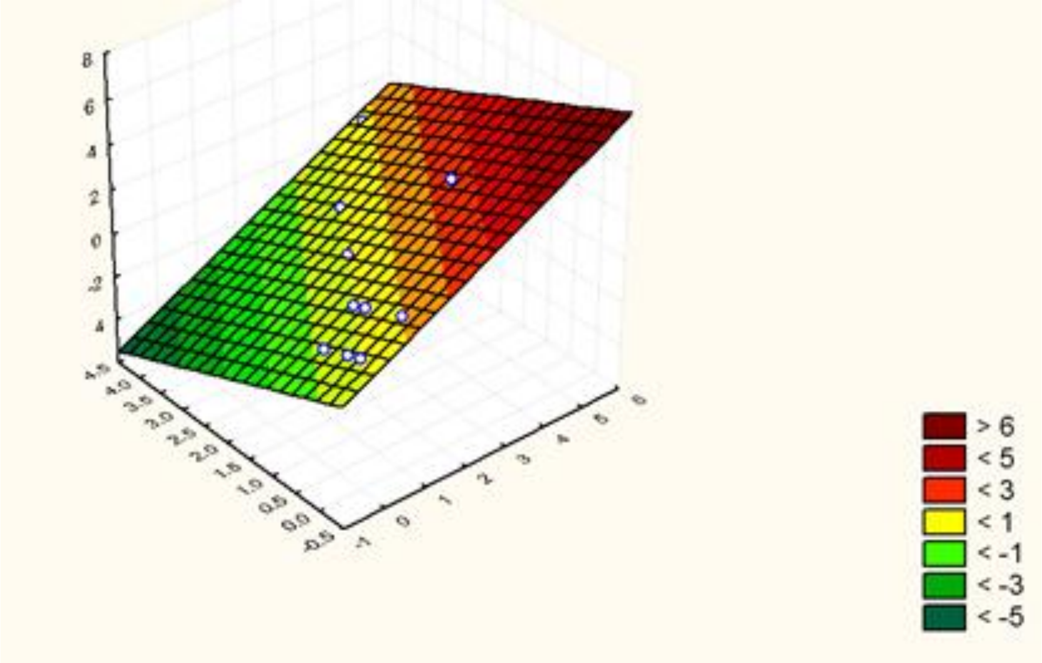

Figure 4. Classification error analysis.

The recommender system is modelled in such a way that on the basis of selected text document, output result lists semantically related teaching materials. This means that, when the student retrieves one lecture material, related teaching materials will be listed as suggestions for those who want to know more about the subject of the lecture.

V. Conclusion and future work

Recently, we witness exponential increase in the amount of information being produced. Effective decision making based on such huge amounts of data can be achieved only if useful knowledge is extracted automatically from them.

The paper described an application of text mining techniques for extraction of useful knowledge from 54 text documents written in Croatian. This required text document pre-processing in order to define key words and form document-term matrix. Based on pre-processing, textual documents have been prepared for *v*-fold cross-validation method in order to define optimal size of document neighbourhood needed to classify it. For classification, kNN algorithm was used.

Applied process was used to classify lecture materials with aim to provide recommendations based on students' interests. This has proven to be useful in multidisciplinary areas such as traffic and transport engineering where students come with different studying backgrounds (e.g. information and communication technologies student needs to understand an multimodal urban traffic simulation teaching example to apply traffic control algorithm with public transport priority for signalised intersections). Application of developed recommender system allows students to consider lectures from different courses (for previous example, lectures on traffic control modelling course), regardless whether they have enrolled for his courses or have not.

In future research, it is planned to include more text documents as input and to create an interface that would allow students to access hyperlinks of suggested teaching materials. This is especially useful for students of multidisciplinary areas and those that have a wish to expand their knowledge.

References

[1] Ò. Celma and P. Lamere, "If you like the beatles you might like...: a tutorial on music recommendation", ACM Multimedia, October 2008, pp. 1157–1158.

[2] A. Rajaraman and J. D. Ullman, "Mining of Massive Datasets", Cambridge University Press, Cambridge, UK, 2011

[3] Y. Wang, J. Hodges, B., Tang, B., "Classification of Web documents using a naive Bayes method", Dept. of Comput. Sci. & Eng., Mississippi State Univ. 2003, pp. 560–565.

[4] R.A Calvo., J. Lee and X. Li, "Managing content with automatic document classification", School of Electrical and Information Engineering, University of Sydney, Australia.2002.

[5] F. T. Hristea, "The Naïve Bayes Model for Unsupervised Word Sense Disambiguation: Aspects Concerning Feature Selection", SpringerBriefs in Statistics, 2013.

[6] J. Y. Pong, R. C. Kwok, R. Y. Lau, Y. Hao and P. C. Wong, "An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata", Journal of Information Science, vol. 34, no. 2, 2008, pp. 213-230.

[7] A.R. Ittoo, Y. Zhang, J. Jiao, "A Text Mining-based Recommendation System for Customer Decision Making in Online Product Customization", The 3rd IEEE International Conference on Management of Innovation and Technology, Singapore, pp. 473-477

[8] S. Loh, F. Lorenzi, R. Saldana, D. Licthnow, "A tourism recommender system based on collaboration and text analysis", Information Technology & Tourism, vol. 6, no. 3, 2003, pp. 157-165

[9] S. Aciar, D. Zhang, S. J. Simoff, J. K. Debenham, "Informed Recommender: Basing Recommendations on Consumer Product Reviews". IEEE Intelligent Systems, 2007, 22(3): pp. 39-47

[10] R.M., Feitosa, S. Labidi, A.L. Silva dos Santos, N. Santos, "Social Recommendation in Location-Based Social Network Using Text Mining", 4th International Conference on Intelligent Systems Modelling & Simulation (ISMS), Bangkok 2013, pp. 67 – 72

[11] S. Venkatraman and S. J. Kamatkar. "Intelligent Information Retrieval and Recommender System Framework", International Journal of Future Computer and Communication, Vol. 2, No. 2, April 2013, pp. 85-89

[12] P. Lops, M. de Gemmis, G. Semeraro, "Content-based Recommender Systems: State of the Art and Trends". Recommender Systems Handbook, 2011, pp.73-105

[13] M. J. Pazzani and D. Billsus, "Content-based recommendation systems. In The adaptive web", Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Lecture Notes In Computer Science, 2007 Vol. 4321. Springer-Verlag, Berlin, Heidelberg pp. 325-341.

[14] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization", Proceedings of the International Joint Conference on Neural Networks, 2003, IEEE, pp. 1661-1666.

[15] M. Hassler. and G. Fliedl, "Text Preparation through Extended Tokenization, Data Mining VII: Data, Text and Web Mining and Their Business Applications". Volume 37. Edited by Zanasi, A and Brebbia, CA and Ebecken, NFF. WIT Press/Computational Mechanics Publications; 2006:pp. 13-21

[16] C. D. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008.

[17] I. Aghayan, N. Noii, M. Metin Kunt, "Extended Traffic Crash Modelling through Precision and Response Time Using Fuzzy Clustering Algorithms Compared with Multi-layer Perceptron", PROMET - Traffic&Transportation, Vol 24, No 6, pp. 455-467

[18] N. Sandhya, Y.Sri Lalitha, A.Govardhan, "Analysis of Similarity Measures for Text Clustering, International Journal of Data Engineering", Vol 2, Issue 4, pp. 1-10

[19] http://www.statsoft.com/, May 2013