Article:
Ševa, J., Schatten, M. and Grd, P. (2015) Open Directory Project based universal taxonomy for Personalization of Online (Re)sources. Expert Systems with Applications, 42 (17-18). pp. 6306-6314. ISSN: 0957-4174

https://doi.org/10.1016/j.eswa.2015.04.033

Open Directory Project based Universal Taxonomy for Personalization of Online (Re)sources

Jurica Ševa*, Markus Schatten, Petra Grd

University of Zagreb, Faculty of Organization and Informatics, Pavlinska 2, 42 000 Varaždin, Croatia

Abstract. Content personalization reflects the ability of content classification into (predefined) thematic units or information domains. Content nodes in a single thematic unit are related to a greater or lesser extent. An existing connection between two available content nodes assumes that the user will be interested in both resources (but not necessarily to the same extent). Such a connection (and its value) can be established through the process of automatic content classification and labeling. One approach for the classification of content nodes is the use of a predefined classification taxonomy. With the help of such classification taxonomy it is possible to automatically classify and label existing content nodes as well as create additional descriptors for future use in content personalization and recommendation systems. For these purposes existing web directories can be used in creating a universal, purely content based, classification taxonomy. This work analyzes Open Directory Project (ODP) web directory and proposes a novel use of its structure and content as the basis for such a classification taxonomy. The goal of a unified classification taxonomy is to allow for content personalization from heterogeneous sources. In this work we focus on the overall quality of ODP as the basis for such a classification taxonomy and the use of its hierarchical structure for automatic labeling. Due to the structure of data in ODP different grouping schemes are devised and tested to find the optimal content and structure combination for a proposed classification taxonomy as well as automatic labeling processes. The results provide an in-depth analysis of ODP and ODP based content classification and automatic labeling models. Although the use of ODP is well documented, this question has not been answered to date.

Keywords: Recommendation systems, content personalization, automatic content classification, automatic content labeling, Information Extraction, Information Retrieval, Open Directory Project, Vector Space Modeling, TF-IDF

1 Introduction

The beginning of the 21st century has witnessed a hyper production of digitally available content. One of the most important processes was made in the redesign of newspapers for the digital generation as they began to present their content online. The

* *Corresponding author*. Tel.: +385 42 390 873; fax: +385 42 213 413. E-mail address: jseva@foi.hr (Jurica Ševa). Address: Pavlinska 2, 42 000 Varaždin, Croatia.

E-mail addresses: jseva@foi.hr (Jurica Ševa), markus.schatten@foi.hr (Markus Schatten), petra.grd@foi.hr (Petra Grd)

downside of this evolution is defined by the paradox of information crisis: the problem of accessing needed information does not lie in the fact that information is inaccessible, but just the opposite; the vast size of digital information users are surrounded with makes it difficult to access appropriate information. One approach in reducing the effects of information crisis is the process of content personalization through recommendation systems. This process can be automated by using automatic content classification and labeling models which is the focus of this work. Automatic content classification has been widely researched and is not a new research field. There are many approaches used in automatic content classification including but not limited to Bayesian classifiers, Support Vector Machines (SVM), Artificial Neural Networks, and clustering techniques (Borges & Lorena, 2010, p. 130). One of the issues with automatic content classification from heterogeneous sources is their different categorization structure. Additionally, there are no universally accepted experimental dataset for large scale hierarchical classification yet, so related work is based on different datasets for evaluation (e.g. ODP, the Yahoo! Directory or some other domain-specific datasets) (He, Jia, Ding, & Han, 2013). Although different datasets are used in different research efforts we can give an overview of weighting schemes used, classification approaches and their results for recent reviewed research efforts that are comparable with this work. We propose the use of ODP[1] Web directory as a unified classification taxonomy. As of time of writing this article an in-depth analysis of ODP and its use as a unified classification taxonomy is not present.

This paper is based on an approach that combines methods and techniques of *information extraction (IE)* (Cowie & Lehnert, 1996), *natural language processing (NLP)*, *information retrieval (IR)* (Salton, 1983; van Rijsbergen, 1979) and *Vector Space Modeling (VSM)* (Salton, Wong, & Yang, 1975) for creating machine understandable classification models used for automatic content classification/labeling. In order to prepare the content of digital textual documents for further processing IE and NLP techniques are used. NLP is a part of Artificial Intelligence (AI) research that allows us to process content presented in natural language and extract tacit knowledge from it. NLP is used to prepare input documents through removing parts of their content that are not useful for further processing. Prepared content is then represented with one of possible weighting schemes. Resulting models and their performance are evaluated based on standard IR measures: *precision (P), recall (R)* and *F1* (all defined below). Python programming language and its extensions *NLTK*[2] (Bird, Klein, & Loper, 2009), *gensim*[3] (Řehůřek & Sojka, 2004) and *scikit-learn*[4] (Pedregosa et al., 2011) have been selected as the implementation platform. NLTK offers a direct way for manipulating human written language and offers a set of tools to prepare the data for further steps and VSM. Genism allows us to represent prepared documents in selected weighting scheme, with *TF-IDF*[5] weighting scheme used in this work. TF-IDF is the oldest and most used weighting scheme in VSM and was initially defined

[1] Open Directory Project
[2] Natural Language Toolkit, http://www.nltk.org
[3] http://radimrehurek.com/gensim/index.html
[4] http://scikit-learn.org/stable/
[5] Term Frequency–Inverse Document Frequency

in (Salton, 1975). It's measure was later expanded upon with idf measure reasoning for which is given in (Robertson, 2004). It is used primarily for VSM, which provides a basis for information retrieval technique(s) used herein. Scikit-learn provides the basis for IR measures implementation and classification model performance.

This paper focuses on testing if ODP presents a good classification scheme for both content-based node classification as well as labeling. We provide several grouping approaches and test optimal number of documents for models in defined grouping schemes for best classification and labeling results. This study aims to meet the following objectives:

(G1) representing ODP content with a set of key words that describe individual nodes based on TF-IDF weighting scheme.

(G2) using ODP structure for automatic classification and labeling based on the content representation defined in *G1*

The rest of the paper is organized as follows: section 2 presents an overview of related work and research efforts this work is based on and compared to. In section 3 an overview of used research methodology is given whilst in section 4 the research results are presented and analyzed. Section 5 concludes on the obtained results, explains the significance of achieved results in the field of intelligent information systems and gives an overview of future work.

2 Related work

The use of folksonomies and/or taxonomies for enhancing information retrieval results is well documented in relevant literature. They are usually used as additional descriptors in various application domains and annotate resources with a defined set of possible labels. In this context there are several web directories available for use in creation of classification taxonomies (*AboutUs.org, Biographicon, LookSmart, Google Directory, Intute, Lycos' TOP 5%, Yahoo! Directory, Zeal etc.*). From all possible and available web directories ODP has been identified as the most suitable for our research agenda due to a number of reasons. ODP itself was the first organized effort to classify Web domains manually into predefined categories and has, from its beginnings, relied on human editors and their manual efforts in classifying submitted Web domains. Therefore it represents an expert-based, pre-labeled collection of documents. The hierarchical structure is presented through 17 root categories and has 0.7 million possible categories (Zhu & Dreher, 2010) with the number of domains listed in the directory exceeding 4.5 million entries. Besides the number of classified web domains, it also presents a hierarchical categorization scheme where each categorized domain belongs to one or more categories that are organized in (maximum) 13 hierarchical levels. All categories are described with one or multiple documents and they represent possible labels in automatic classification/labelling system.

One of the main problems in using existing taxonomies (e.g. ODP) is the structure of data presented in the taxonomy and its combination in created classification models. The majority of research efforts try to utilize preexisting connections and hierarchical structure from each specific data source used in automatic classification re-

search efforts. In case of ODP, classification models can be created based on different grouping schemes as presented in this work. Documents used in classification models can be grouped based on parent-child relations or sibling relations. Additionally, an alternative way of grouping data is via symbolic links that are present in most predefined web directories. A symbolic link is a hyperlink which makes a directed connection from a webpage along one path through a directory to a page along another path (Perugini, 2008). As their results show, almost 97% of symbolic links results with multiclassification and *"majority of symbolic links (>77%) are multiclassification links which connect two categories which share at least the first two levels of topic specificity*" (Perugini, 2008, p. 927). The majority of symbolic links produce multiclassification this approach will not be used as their use generates additional noise in the classification and labeling process. Additionally, reviewed research efforts differ based on VSM weighting scheme used (mostly TF-IDF) as well as the range of the taxonomy used (domain-specific branches or the entire taxonomy). The majority of research efforts that use ODP for automatic classification are domain-specific, use TF-IDF weighting scheme and limit the number of ODP documents, both in hierarchical branches as well as hierarchical depth, used in created classification models.

Marath, Shepherd, Milios, & Duffy (2014) focus on the Yahoo! Directory and present a unified classification model or framework for highly imbalanced hierarchical datasets. In their work ODP was used as the validation data set. They focus on a subset of ODP and use 17,217 categories and 130,594 web pages from ODP data whilst we focus on the entire directory. Additionally, their work uses standard machine learning algorithms for classification whilst we focus on VSM based models and IR. Classification results are evaluated using F1 measure and as reported they achieve macro-averaged F1-Measure of the DMOZ subset of value 84.85%. ODP is used as the testing set again in (Rajalakshmi & Aravindan, 2013). This approach uses just the URL of a document for its classification but they use 3-gram notation for feature extraction whilst we use 1-gram notation. Classification models are built with SVM and Maximum Entropy classifier. Their testing set was again limited, this time to 14 root categories, and Fl was used as the evaluation metric, with classification results around 80% for each of selected root categories, which is lower than our results. Zubiaga & Ji (2013) use ODP for the classification of data available over Twitter. He et al. (2013) focus on hierarchical classification of rare categories in ODP. They propose an approach based on LDA[6] (Blei, Ng, & Jordan, 2003). Their classification models are created by SVM and use term frequency vectors for document representation. Their experiments where performed on Chinese Simplified branch of the DMOZ directory which has 13 root categories and a hierarchical depth of 6. Again, we use a larger part of ODP data in our classification models. As evaluation measures standard P, R and F1 measures were used. Their overall classification results based on their approach is below 80% for all proposed classification schemes. Amini, Ibrahim, Othman, & Nematbakhsh (2015) use ODP in combination with other web directories for a reference ontology in the scope of scientific publishing. From all available categories in

[6] Latent Dirichlet Allocation

ODP, they focus on the Computer Science section of the directory leaving them 8471 general entries.

Fathy, Gharib, Badr, Mashat, & Abraham (2014) use ODP for improving search results based on user preferences. ODP and its concepts are used as additional descriptors for user search queries. Reference taxonomy, based on TF-IDF weighting scheme, chooses the first 30 URLs for each concept based on the order in which they are represented by ODP. ODP is additionally used for construction user profiles where search results clicked by the user are classified into concepts from ODP which are then used together to build the profile. Duong, Uddin, & Nguyen (2013) also focus their research efforts on enhancing search results by using ODP as the basis for a reference ontology used to additionally label visited documents. Again, documents in ODP were represented with TF-IDF weighting scheme based vectors. These vectors are then used to search for similar ontological concepts. Their research is focused on user searches in academic domain of computer science and therefore their models only include that branch of ODP. Their experimental data set consists of 650 concepts and 15,326 documents that were indexed under various concepts. Results were evaluated on P, R and F1 measures although results are only presented graphically.

In (Lee, Ha, Jung, & Lee, 2013) ODP was used as an additional descriptor in the domain of contextual advertising. They prune down ODP data used for training and testing down to 15 root categories, 95,259 domains, 5,178 nodes and a maximum of nine levels that are used to create the taxonomy. Documents are represented based on TF-IDF weighting scheme values. Their results are evaluated based on P, R and F1 with best P results at 0.863. Vargiu, Giuliani, & Armano (2013) also focus on contextual advertising and use collaborative filtering for classification models creation. It uses ODP and its data to classify the page content and to suggest suitable ads accordingly. The use TF-IDF weighting scheme to transform prepared documents for classification. They use Rocchio classifier to created centroids and classify the document in to one or more ODP categories.

Two recent research efforts in were based on the entire ODP dataset. Yun, Jing, Yu, & Huang (2012) focus on combining data from ODP and Wikipedia where ODP is used to define a set of terms that are then compared with Wikipedia concepts. Their work is combined in Two-level Representation Model (2RA) and uses syntactic information and semantic information extracted from Wikipedia data. Term-based VSM and TF-IDF weighting scheme are used in syntactic level to record the syntactic information. Semantic level consists of Wikipedia concepts related to the terms in the syntactic level. Their classification approach, defined with Multi-layer classification (MLCLA) framework, is designed to handle large scale data with complex and high dimensions layer-by-layer. Their best achieved classification results, measure with F-score measure, differ for SVM classification (0.9942) and 1NN classification algorithms (0.8468). Ha, Lee, Jang, Lee, & Lee (2014) focus on using various classification algorithms for text classification and conclude that training data expansion significantly improves the classification performance. They focus their research efforts on the best approach of hierarchically pruning the ODP tree while traversing available branches from root node towards deeper hierarchical levels. They also remove two categories (Regional and World respectively) from training and testing data which

leaves them with 182,003 categories and 1,228,843 web pages. As the weighting scheme they also utilize TF-IDF weights and base their classification approach on generated merge-document and merge-centroid vectors. They measured the accuracy of a classifier as the number of correctly classified test data divided by total number of test data, based on two F-measure values (macroaveraged (maF_1) and micro-averaged (miF_1) F-measure). Although they give a comparison of different classification algorithms used, their best classification results yields at approximately 36%.

Compared to presented approaches in reviewed literature we use ODP purely as the basis for a universal classification taxonomy. The focus of our approach is to enable personalization of news articles from various online news portals. Due to their heterogeneous classification scheme a universal classification scheme is needed to provide a general classification scheme. For these purposes we analyze the entire ODP content and don't exclude categories either based on their depth or the number of documents describing the category. Although ODP is used in different application domains such an approach is not currently presented in recent research efforts. Our work also uses specific steps in preparing ODP data for classification models by utilizing NLP and IE tools and techniques for dimension reduction. Additionally, we propose a two-step classification approach where the first stage is focused on general classification and second stage attaches multiple labels to the classified resource. For these purposes we show different approaches in grouping ODP content and compare their evaluation results. Compared to presented relevant research efforts, our approach performs as good or better.

3 Methodology

IE was defined and first presented in (Cowie & Lehnert, 1996) with its goal defined as "*creating a system that finds and links relevant information while ignoring extraneous and irrelevant information*". IR "*deals with the representation, storage, organization of, and access to information items*" (Baeza-Yates & Ribeiro-Neto, 1999). It was presented as a topic in the early 1950's with the emergence of computers and its scope has increased in 1970's through the work of Van Rijsbergen (van Rijsbergen, 1979) and Salton (Salton, 1983). Salton also presented the foundations of VSM approach for document modeling in (Salton et al., 1975). VSM in general, as a model for IR, is first proposed in (Salton, 1979).

TF-IDF is a combination of two measures describing a document compared to a document collection (classification model): TF (term frequency) and IDF (inverse document frequency). The weighting scheme is then defined as

$$\text{TF-IDF}(t, d, N) = \text{tf}(t, d) * \text{idf}(t, N), \quad (1)$$

with

$$\text{tf}(t,d) = t_d/d_t \quad (2)$$

and

$$\mathrm{idf_t} = \log\left(\mathrm{N} / \mathrm{df_t}\right) \tag{3}$$

where t is the observed expression, d is a document from the collection of N documents, t_d is the number of times term t appears in a single document, d_t is the total number of terms in the document and df_t is the number of documents from N containing term t. This measure assigns a value to the observed expression t in document d that is:

- greatest where t is common in a small number of documents,
- smaller when t is less common in d, or when it appears in many documents,
- smallest when t appears in all documents in N.

As stated in (Yun et al., 2012) "*VSM is the most popular document representation model for text clustering, classification and information retrieval*". Set of measures for IR model evaluation, with precision (P), recall (R) and F1 measures used in this work, were first presented in (Salton & Lesk, 1968). *Precision (P)* is defined as the fraction of retrieved documents that are relevant:

$$Precision = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)} = P(relevant|retrieved) \tag{4}$$

Recall (R) is defined as the fraction of relevant documents that are retrieved:

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)} = P(retrieved|relevant) \tag{5}$$

F-measure is defined as the weighted harmonic mean, known as F1, of P and R:

$$F1 = 2 * P * R / (P + R) \tag{6}$$

ODP and its content and structure data files are freely available on the ODP Web page[7] in RDF[8] format. For a detailed presentation of the data available in ODP RDF dump files see (Kalinov, Stantic, & Sattar, 2010). Due to its structure ODP data has to be grouped together in order to create useful classification models. In this work there are several grouping schemes devised, as presented in 3.3.

The reason for different comparison models is to determine the following:

1. *overall quality of the proposed universal taxonomy for automatic document classification via ODP-based comparison models*
2. *optimal grouping scheme and model size for future use, both for classification and automatic labeling*

NLTK framework is a platform that offers interfaces for corpora[9] and lexical resources like WordNet (Miller, 1995) which makes it easier to implement needed natural language processing tasks as explained in (Perkins, 2010). This framework, in this work, has been used for data cleaning purposes and removing all textual data that

[7] http://rdf.dmoz.org/

[8] Resource Description Framework

[9] Collection of 'real word' texts used in NLP analysis

didn't have any value for further analysis (e.g. HTML tags, stop words, first/last names, grammatical POS[10] parts of text). Additionally NLTK was also used for stemming with Porters stemming algorithm (Porter, 2006). Stemming "*is designed to remove and replace well known suffixes of English words*" (Perkins, 2010, p. 26) thus giving us the root form of selected word. This way document content normalization can be achieved.

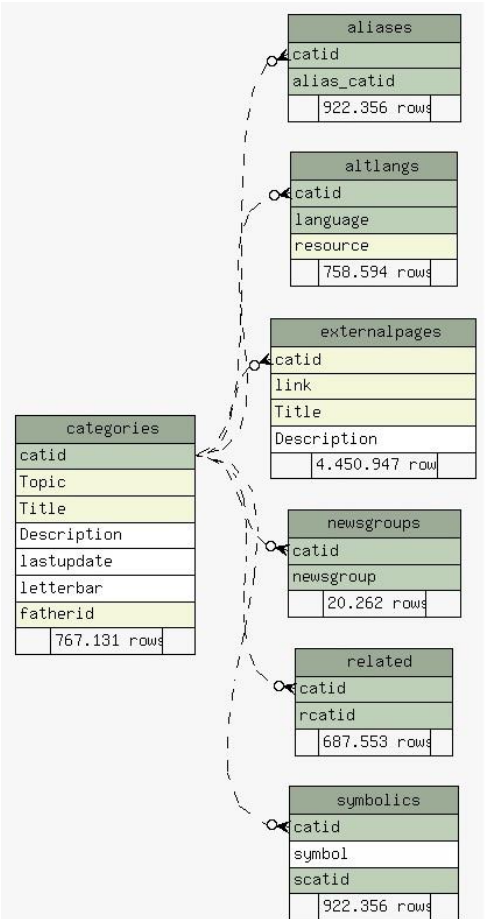

Fig. 1. Open Directory Project MySQL structure

ODP data was extracted and stored in a MySQL database with help of the open source tool *suckdmoz*[11]. The database scheme created by this tool is presented in Fig. 1. Two created database tables are especially interesting for further analysis: *'dmoz_categories'* and *'dmoz_externalpages'*. They offer a list of classified domains

[10] Part of speech

[11] http://sourceforge.net/projects/suckdmoz/

available in ODP along with their respective descriptions. These descriptions are the basis for crated classification and labeling models. The overall process of web usage mining is presented in Fig. 2. Steps specific for this research, with the goal of creating ODP-based universal classification models which will be described in more detail in the following subsections, are as follows:

1. data preparation
2. indexing
3. similarity evaluation
4. model evaluation

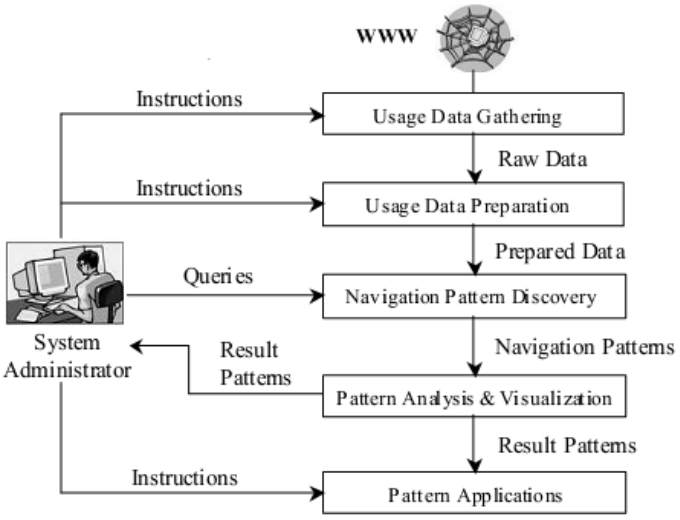

Fig. 2. Generalized Web usage mining system (Hu, Zong, Lee, & Yeh, 2003)

3.1 Data preparation

Raw data, available through ODP database dump, has been prepared for further data analysis. During the preprocessing phase firstly two categories were removed from the ODP data dump. ODP branches for root categories 'Adult' and 'World' were excluded from further analysis due to their content either not being written in English or being multimedial data (e.g. digital images). After that 15 root categories remained.

Afterwards, hierarchical depth levels were defined based on two approaches: (1) URL-based classification scheme descriptor with delimiter '/', and (2) bottom-up approach, based on parent-child relationship, using the '*fatherid*' column, where each document on level *n* is described with both '*fatherid*' and '*catid*' values (as shown in Fig. 1). In this approach '*fatherid'* on level *n* references '*catid*' value on level *n – 1*. Depth information is stored in the column '*depthCategory*'.

Finally, entries with an empty '*Description*' column in tables '*dmoz_categories*' as well as '*dmoz_externalpages*' were assigned a special value '*-1*' in column '*filterOut*'. This value marked all database rows that were excluded from both training and testing data.
Using the above mentioned filtering steps available data in tables *'dmoz_categories'* and *'dmoz_externalpages'* was reduced as shown in Table 1.

Table 1. Available data after filtering

Database table	*Original rows*	*Prepared rows*	*% of rows left*
'dmoz_externalpages'	4 592 105	2 637 412	~57%
'*dmoz_categories"*	763 378	496 007	~65%

3.2 Indexing

Indexing is focused on extracting text features. The algorithm for this process, which is a modified version from (Greenwood, 2001), reads as follows:

1. While there are documents load the next document
2. Split the document into tokens in 1-gram notation (defined by a predefined delimiter)
3. Remove:
    a. HTML element tags (e.g. *<HEAD>, <BODY>, <DIV> etc.*) and special formatting HTML tags (e.g. *<b>, <i> etc.*)
    b. punctuation signs
    c. known male/female first names
    d. single alphanumeric characters
    e. stop words (two stop word lists were used; NLTK based list as well as manually created list)
4. Stem resulting tokens
5. If there are more documents, go to 1.

The overall reduction of the number of words is approximately 47%, which shows that by using the steps in the presented algorithm one can achieve a significant dimension reduction for further analysis.

3.3 Similarity evaluation

The prepared data is represented with TF-IDF weighting scheme and serves as input for the classification models. Two main approaches for creation and testing of prepared models have been devised. Each model is defined through used ODP content grouping scheme and number of documents in created model. Results and their evaluation are shown in the next section.

Available content for classified web domains is first grouped together based on either '*catid*' or '*fatherid*' column values and are as follows:

- *GENERAL grouping*, where a single document in a category model is represented by a single document from a specific category
- *CATID grouping*, where a single document in a category model is represented by all documents with the same *'catid'* value from a specific category
- *FATHERID grouping*, where a single document in a category model is represented by all documents with the same '*fatherid'* value from a specific category

Next, for each grouping scheme two main size model families were created:

1. *Percentage models*, where, for each of the main 15 categories, first 25, 50, 75 and 100% of documents were used in model creation:
2. *Limit models* where, for each of the main 15 categories, first 1000, 2500, 5000, 7500, 10000 and 20000 documents were used in model creation

The purpose of different grouping and model document number schemes is to test:

1. if ODP is a good source for the proposed universal classification taxonomy and
2. if there are differences in evaluated IR measures for different model creation approaches related to different grouping schemes and number of documents used.

The model creation process was as follows:

1. prepare input data, following steps from sections 3.1 and 3.2
2. create dictionary, with the list of all tokens/words taken from the database for each specific category
3. create corpora
4. create VSM representation based of TF-IDF weighting values

The difference between models is defined in the first step. Files, created as the result of this stage, are then used in testing and model evaluation.

3.4 Model evaluation

The data available from ODP was divided in two distinct sets, training set and testing set, with their ratio being 80/20. Achieved results were evaluated with standard IR measures *P, R* and *F1*. Evaluation results answered two research question defined in section 1. The results of the evaluations were stored in a MySQL database for further analysis.

The overall steps for model evaluation where the same for both research questions and re as follows:

1. Get *n* sample documents from testing data set
2. For each sample document:
    a. Prepare the sampled documents (following the steps described in sections 3.1 and 3.2)
    b. Load comparison model file[12]
    c. Calculate similarity value of each sample document against loaded comparison model with the following constraints;

[12] Gensim generated TF-IDF weighting scheme file

b) Rank documents by similarity value (descending)
c) Filter out documents with similarity value below set minimum similarity value[13] (limited to 1000 most similar documents)

3. Evaluate results

4 Evaluation Results

Results evaluation is focused on answering two research questions:

1. *Overall classification quality of ODP* by comparing training set models from category X against testing set data for all categories
2. *Best grouping scheme for automatic labeling* by comparing training set models from category X against testing set data for category X for each grouping scheme

The first research question is focused on determining if ODP is a suitable candidate for content classification of unclassified documents. The results of this process are of vital importance for the rest of research agenda. Furthermore, due to multiple possibilities of combining ODP data multiple grouping schemes were devised. Hence, second research question was devised and tested to show which grouping scheme yields best classification results.

Data available in ODP, prepared as explained in sections 3.1 and 3.2, was divided in two document sets: training document set, used to create classification models, and testing document set, used to test classification models. A requirement was set for both data sets: they should have at least one document each with the same *'catid'* and *'fatherid*' values. In both approaches the evaluation was done by comparing the classification models on the document training set data and comparing *'catid'* and/or *'fatherid"* values, depending on the grouping scheme tested, of the input documents and the returned documents sorted by descending similarity value. The results were evaluated with standardized IR measures: *P*, *R* and *F1*. Next, a detailed presentation and explanation of evaluation results is provided.

4.1 Overall classification quality of ODP

First we determined whether ODP can be used as a classification taxonomy at all. For these purposes a simple testing scheme was derived and implemented where, based on *n* documents form category X, a set of documents from the testing set was evaluated against every created model for each of the proposed grouping schemes (*GENERAL*, *CATID* and *FATHERID*).

Calculated similarities for tested documents against different grouping scheme models where summed for each compared category. Stored data tested which category, based on the overall sum of all returned similarity values, had the highest cumula-

[13] Documents in comparison model whose similarity to the analyzed document is below set similarity value

tive similarity value; that category was shown as the most similar one from all 15 possible categories in comparison to testing data from category X.

Table 2. Overall classification results of root categories

Grouping scheme / classification category	***CATID***	***FATHERID***	***GENERAL***
Positive	4.4	8.5	8.3
Negative	10.6	6.5	6.7
Number of categories	*15*	*15*	*15*
% positive classification	*30*	*56*	*55*

The overview results are shown in Fig. 3. When it comes to the proposed grouping schemes, the grouping based on *CATID* (positive with value 4.4/15 and negative with value 10.6/15) showed the worst results based on cumulative similarity value. This is to be interpreted as follows: from all testing data document only approximately 30% were classified in to their original category. The devised classification scheme performs better for other two proposed grouping schemes as shown in Table 2.

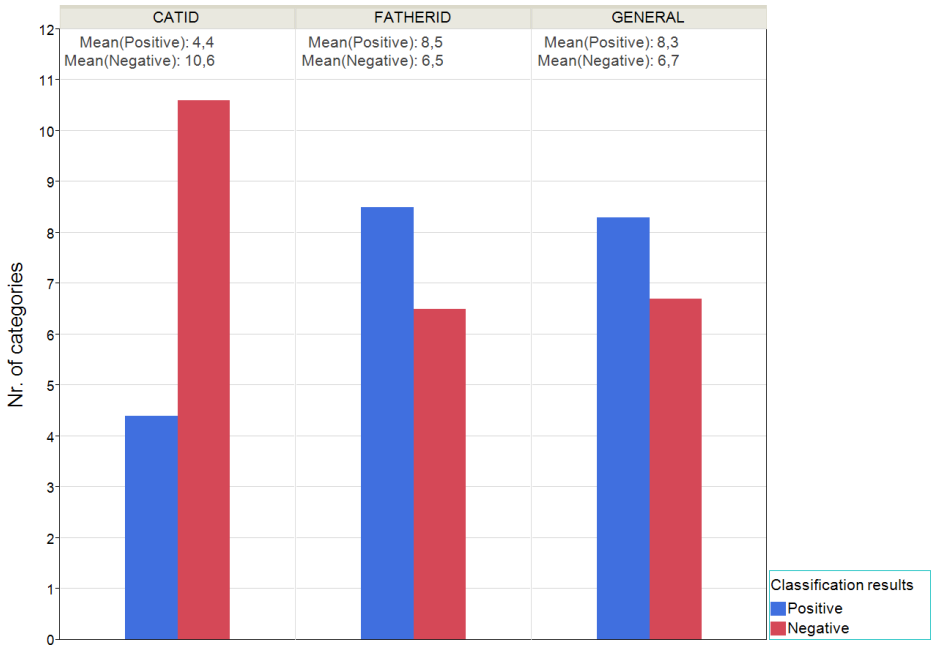

Fig. 3. Overall classification quality of ODP

This shows the potential of using ODP as a universal taxonomy and suggests that the classification quality directly depends on how the data is prepared and grouped together.

Although these results can be interpreted as not sufficient when we provide additional constraints for each grouping scheme and limit the number of documents included in testing models we get a better overview of the nature of ODP and its data. This overview suggests that ODP based classification models provide a good basis for overall content classification when limiting the number of documents included in classification models. For all three grouping schemes several model size families have been devised to test models with different document numbers. The results are shown in Fig. 4 and presented in detail in Table 3.

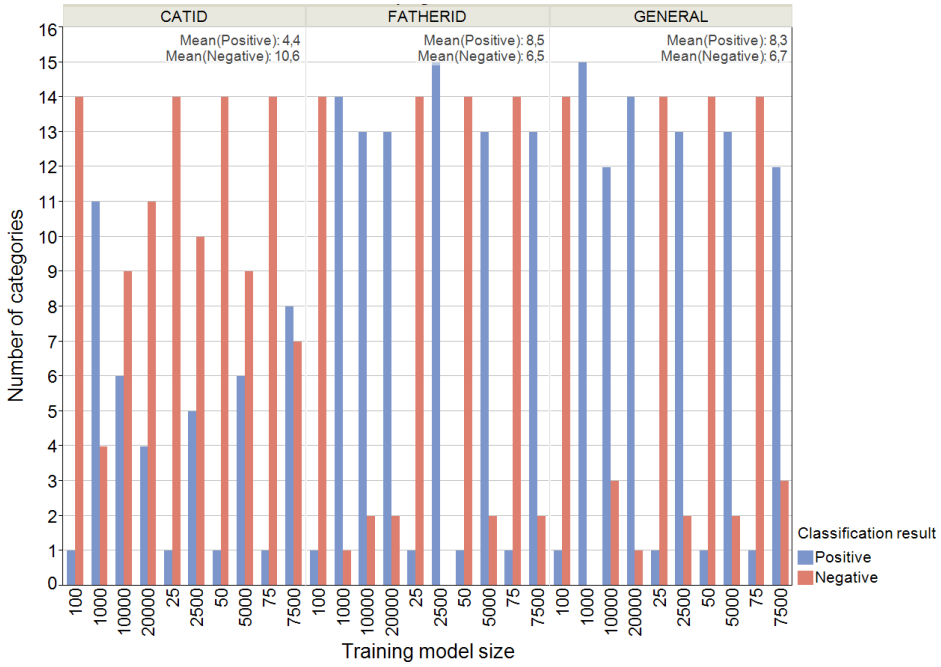

Fig. 4. Overall classification quality of ODP with different grouping schemes

When it comes to number of documents used in generated classification models evaluation results show that *percentage-based models* are behaving subpar and actually, due to the different number of documents they are made of, increase the amount of noise in the created models. *Limit based models* provide far better results and their use in future research is suggested by these results. As far as used grouping scheme is concerned, *CATID* grouping scheme yields the worst results once again, but this time independently to the number of documents used in classification models. *FATHERID* and *GENERAL* grouping schemes perform below par when used in combination with percentage models but yield better results when used in combination with limit models. Additionally, as the number of documents used in classification models increases classification results worsen. It is easy to deduce that the grouping scheme is not the only factor in achieving good classification results but is additionally improved when limiting number of documents used for classification models. Our evaluation results

suggest that smallest classification models are to be used for overall classification as they include enough information for good overall classification with *GENERAL* grouping scheme models providing best results (100%).

Table 3. Overall classification results of root categories with different grouping schemes

GROUPING		*CATID*		*FATHERID*		*GENERAL*	
	RESULT	*Positive*	*Negative*	*Positive*	*Negative*	*Positive*	*Negative*
Percentage models	*25*	1	14	1	14	1	14
	50	1	14	1	14	1	14
	75	1	14	1	14	1	14
	100	1	14	1	14	1	14
Limit models	*1000*	11	4	14	1	15	0
	2500	5	10	15	0	13	2
	5000	6	9	13	2	13	2
	7500	8	7	13	2	12	3
	10000	6	9	13	2	12	3
	20000	4	11	13	2	14	1

4.2 Choosing best grouping scheme for automatic labeling

The goal of this process is to finely tune the classification and to apply automatic labels to the active document, as well as to test the quality of ODP as a possible labeling scheme. In this step only labels from the most similar category, as classified in previous section, are used. Evaluation is based on the same steps and data as used in previous section. The IR measures used are calculated on the ratio between tested document(s) (*input value*) and returned most similar document (*output value*). Compared values are either for database fields *'catid'* or *'fatherid'*, depending on the grouping scheme used (*CATID* and *FATHERID* respectively). Evaluation results are presented and discussed next.

Table 4. Best grouping scheme evaluation for automatic labeling (overall)

Grouping scheme / IR measures	***CATID***	***FATHERID***	***GENERAL***
Precision	0,92617	0,904447	0,92037
Recall	0,60799	0,21546	0,91859
F1	0,70016	0,31114	0,91651

First we determine which of the proposed grouping schemes is best used in the process of automatic labeling. Overall labeling results, for different grouping schemes, are shown in Table 4. The results show that, grouping scheme wise, the differences between different grouping schemes are small but only for the *P* value. The results for other two measures, *Recall* and *F1*, indicate that the *GENERAL* group-

ing scheme based models are to be used for automatic labeling. They are followed by *CATID* and *FATHERID* grouping scheme based models.

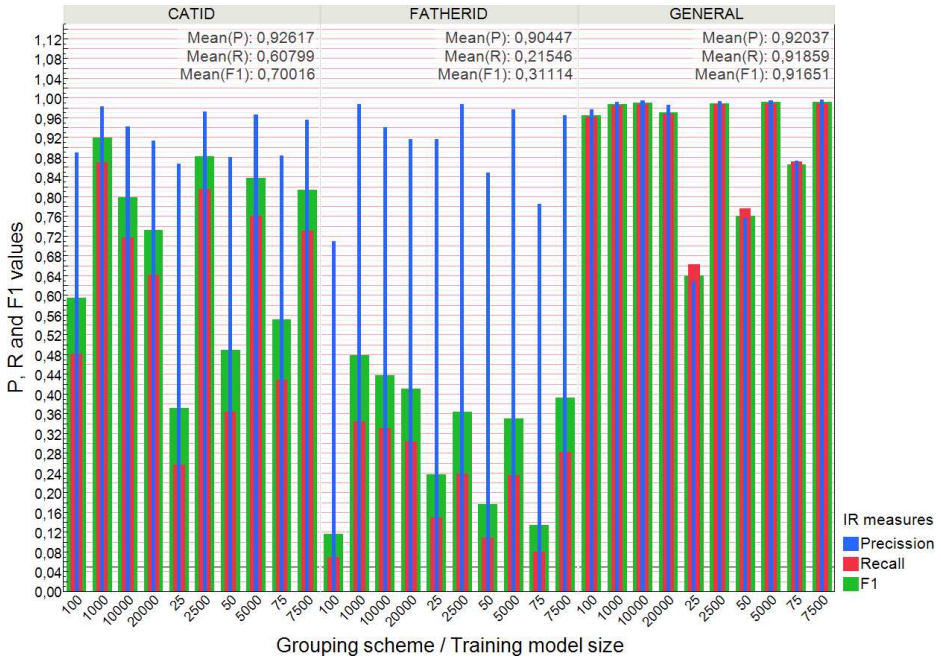

Fig. 5. Best grouping scheme evaluation

A more detailed look gives us a better insight in to the suggested scheme. As far as the number of model documents to be used for automatic labeling is concerned, as presented in Fig. 5 and shown in detail in Table 5, the proposed *Percentage models* are returning poor results as far as R and F1 measures are concerned, while P measure results are satisfactory. The results are conclusive across all three grouping schemes and the best performing percentage model (*100% percentage model*) is performing worse than the worst performing limit model (*20 000 limit documents*).

Limit models proved to yield better results for automatic labeling. Differences between different limit models fluctuate. The best limit model results, as shown in Table 5, are given for limit model *7500* and *GENERAL* grouping scheme with all other limit models performing better than percentage models. The results are promising as far as the process of automatic labeling is concerned. These results will be used in future research dealing with ODP-based content labeling.

Table 5. Best grouping scheme for automatic labeling detailed results

Grouping		CATID			FATHERID			GENERAL		
Model size	IR measure	P	R	F1	P	R	F1	P	R	F1
Percentage models	25	0,867	0,258	0,372	0,918	0,151	0,238	0,906	0,914	0,908
	50	0,881	0,366	0,491	0,850	0,110	0,178	0,758	0,777	0,763
	75	0,885	0,431	0,553	0,787	0,082	0,136	0,874	0,872	0,866
	100	0,891	0,482	0,596	0,711	0,070	0,117	0,978	0,962	0,966
Limit models	1000	0,984	0,871	0,920	0,988	0,345	0,480	0,993	0,986	0,989
	2500	0,973	0,817	0,883	0,988	0,239	0,365	0,995	0,988	0,990
	5000	0,967	0,762	0,840	0,978	0,237	0,352	0,996	0,990	0,993
	7500	0,956	0,732	0,815	0,966	0,283	0,395	0,997	0,991	0,993
	10000	0,944	0,718	0,799	0,942	0,331	0,440	0,996	0,988	0,991
	20000	0,915	0,643	0,733	0,918	0,305	0,411	0,987	0,968	0,973

5 Result analysis and future work

The objectives of this study were to test ODP as the proposed universal taxonomy for classification and automatic labeling. Such a taxonomy is used in the domain of recommender systems when dealing with multiple sources, each with its own information structure. Such a classification can be achieved in two steps; first, unclassified document is classified in one of 15 root categories identified in ODP and secondly additional categorization labels are attached to the analyzed document. Due to the structure of ODP there are several possibilities for organizing its content for classification and labeling models. Our research presents an in-depth look in to the best way of grouping ODP data together and optimal number of documents in created classification models. For these purposes three grouping scheme and two model size families have been devised. Based on evaluation results, best grouping and size models have been identified both for classification as well as labeling steps. Such an extensive ODP analysis has not been found in the reviewed literature.

When it comes to the classification step first the overall adequacy of ODP as the proposed taxonomy was evaluated. Evaluation tested if an original document will be classified in the originating category or if it will be classified as a member of an alternative category. Possible categories were 15 root categories left after ODP data preparation. The results show that, as far as the overall classification quality of ODP is concerned, evaluation results depend on the used grouping scheme and additionally to the number of documents used in classification model. Best results are achieved when using *GENERAL* grouping scheme model based on 1000 documents from ODP. When it comes to the second step, automatic labeling, same labeling models regarding grouping scheme and size limits are used. The purpose of this evaluation is to determine the best grouping scheme and size limitation combination for automatic label-

ing. The results show that limit models perform better than percentage models in all cases and that the best performing model is based on *GENERAL* grouping scheme and 7500 documents. This is expected as percentage models take different number of documents in consideration while creating labeling models. Limit models on the other hand use the same number of documents for labeling models. When compared with related and reviewed work, our classification approach and created models achieve better results both for overall classification (with our best performing classification model achieving 100% precision) as well as automatic labeling (99,7). We have to stress out that these results are achieved when evaluating on ODP testing data.

Although current results are satisfactory for two of three proposed grouping schemes there is room for improvement by using additional content preparation techniques (e.g. n-gram notation with n > 1). Additionally, we can, based on achieved results deduce that frequency based analysis is not the best approach for this web directory. Besides additional steps in preparing ODP's content LDA can be used as the basis for classification models. Next to IR we can also test different machine learning techniques as used in the several reviewed articles. One analysis that is missing is algorithm performance in terms of speed of execution. Although our results are satisfactory there are several news taxonomies that one can use in addition to ODP such as Wikidata[14], DBpedia[15] ontology and other dictionaries to create better performing models when classifying actual news items. The results of this research are implemented as part of the system RecommendMe[16].

6 References

Amini, B., Ibrahim, R., Othman, M. S., & Nematbakhsh, M. A. (2015). A reference ontology for profiling scholar's background knowledge in recommender systems. *Expert Systems with Applications*, *42*(2), 913–928.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison Wesley.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Borges, H., & Lorena, A. (2010). A Survey on Recommender Systems for News Data. *Smart Information and Knowledge Management*, 129–151.

[14] https://www.wikidata.org/
[15] http://dbpedia.org/
[16] http:///rec.foi.hr:5000

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, *39*(1), 80–91.

Duong, T. H., Uddin, M. N., & Nguyen, C. D. (2013). Personalized semantic search using ODP: A study case in academic domain. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7975 LNCS*(PART 5), 607–619.

Fathy, N., Gharib, T. F., Badr, N., Mashat, A. S., & Abraham, A. (2014). A Personalized Approach for Re-ranking Search Results Using User Preferences. *Journal of Universal Computer Science*, *20*(9), 1232–1258.

Greenwood, M. (2001). Implementing a Vector Space Document Retrieval System. *Dcs.shef.ac.uk*.

Ha, J., Lee, J.-H., Jang, W., Lee, Y.-K., & Lee, S. (2014). Toward robust classification using the Open Directory Project. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 607–612).

He, L., Jia, Y., Ding, Z., & Han, W. (2013). Hierarchical classification with a topic taxonomy via LDA. *International Journal of Machine Learning and Cybernetics*, *5*(4), 491–497.

Hu, C., Zong, X., Lee, C., & Yeh, J. (2003). World Wide Web Usage Mining Systems and Technologies. *SYSTEMICS, CYBERNETICS AND INFORMATICS*, *1*(4), 53–59.

Kalinov, P., Stantic, B., & Sattar, A. (2010). Building a Dynamic Classifier for Large Text Data Collections. In H. T. Shen & A. Bouguettaya (Eds.), *Proceedings of the Twenty-First Australasian Conference on Database Technologies - Volume 104 (ADC '10)* (Vol. 104, pp. 113–122). Australian Computer Society, Inc., Darlinghurst, Australia.

Lee, J.-H., Ha, J., Jung, J.-Y., & Lee, S. (2013). Semantic contextual advertising based on the open directory project. *ACM Transactions on the Web*, *7*(4), 24:1–24:22.

Marath, S. T., Shepherd, M., Milios, E., & Duffy, J. (2014). Large-Scale Web Page Classification. *2014 47th Hawaii International Conference on System Sciences*, 1813–1822.

Miller, G. a. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perkins, J. (2010). *Python Text Processing with NLTK 2.0 Cookbook* (First edit.). Packt Publishing.

Perugini, S. (2008). Symbolic links in the Open Directory Project. *Information Processing & Management*, *44*(2), 910–930.

Porter, M. F. (2006). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *40*(3), 211–218.

Rajalakshmi, R., & Aravindan, C. (2013). Web page classification using n-gram based URL features. In *2013 Fifth International Conference on Advanced Computing (ICoAC)* (Vol. 263, pp. 15–21). IEEE.

Řehůřek, R., & Sojka, P. (2004). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges* (pp. 45–50). ELRA.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520.

Salton, G. (1975). A Theory of Indexing. *Regional Conference Series in Applied Mathematics*.

Salton, G. (1979). Mathematics and information retrieval. *Journal of Documentation*, *35*(1), 1 – 29.

Salton, G. (1983). *Introduction to Modern Information Retrieval*. Mcgraw-Hill College.

Salton, G., & Lesk, M. E. . (1968). Computer Evaluation of Indexing and Text Processing. *Journal of the ACM*, *15*(1), 8–36.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann Newton, MA, USA.

Vargiu, E., Giuliani, A., & Armano, G. (2013). Improving Contextual Advertising by Adopting Collaborative Filtering. *ACM Transactions on the Web (TWEB)*, *7*(3), 1–22.

Yun, J., Jing, L., Yu, J., & Huang, H. (2012). A multi-layer text classification framework based on two-level representation model. *Expert Systems with Applications*, *39*(2), 2035–2046.

Zhu, D., & Dreher, H. (2010). Characteristics and uses of labeled datasets - ODP case study. In *Proceedings - 6th International Conference on Semantics, Knowledge and Grid, SKG 2010* (pp. 227–234). Ieee.

Zubiaga, A., & Ji, H. (2013). Harnessing Web Page Directories for Large-Scale Classification of Tweets. In *WWW '13 Companion Proceedings of the 22nd international conference on World Wide Web companion* (pp. 225–226).