Clustering on Social Web

Tomás Kuzár[*]
Institute of Applied Informatics
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
tomas.kuzar@gmail.com

Abstract

Social web increases its potential rapidly. Growing number of involved users leads to significant increase in amount of user-generated content. End users have great opportunity to express themselves by publishing statuses, blogs or photos and in the meantime they consume the content generated by others. In our work we focus on process of social web data consumption - gathering, processing and visualization. In our research we focus on processing of unstructured textual content of Social Web in order to achieve more efficient access to relevant information. We have designed and evaluated methods for building precise content clusters by mining social web data. Our findings indicate the need to encounter external knowledge and the internal relationships between objects on social web to increase the accuracy of extracted knowledge. In user study we demonstrate how the accurate content clusters augment the access to relevant information on the social web.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*; I.5.4 [**Pattern Recognition**]: Applications—*Text processing*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

Keywords

Social Web, Blogs, Content Clustering

1. Introduction

Web 2.0 empowered broad audience of web users to participate on web content creation. Everyone got a chance to be a data producer. In meantime emergence of the Web 2.0 created a group of innovative applications like blogs, wikis, social networking sites, photo sharing applications, content bookmarking tools and many others to be used by very broad audience. Social web can be understood as huge and valuable source of information. But as amount of data and information is enormous, there are many attempts to serve the information in filtered, aggregated way to the data consumer.

On the other hand the companies use Social web as valuable source of information and also as a channel to promote their products to prospective customers. In the same time they use the knowledge extracted for the social web to support their decision making.

Huge amount of this information can be processed automatically and that can augment the user browsing and searching experience. In past decades there was a significant increase of usage of software applications in order to support business processes. With emerge of Web 2.0 many existing software applications started to move to the web environment.

In the past there was a challenge to find some information. Nowadays there is a challenge to pick relevant information and put it into right context. Especially business data analysts in role of data consumers ask: How could we benefit from the potential of social web data? Social web increases its power very fast. It consists of huge amount of user generated content which is spread across the web. This content is already being used for different purposes - fun, brand building, marketing, sales support or information search. There is a high number of methods and algorithms available which can process enormous amount of information available on Social Web. But there methods still can not uncover potencial hidden in Social Web. The methods are still not precise enough and the potential of the potential hidden in that enoumous amount of data is still not fully discovered.

We recognized several open questions related to access to information on Social web:

- Adoption of known methods and approaches to Social Web environment
- Design of new methods specific for Social Web environment
- Processing of massive datasets
- Social Web data visualization

[*] Recommended by thesis supervisor: Prof. Pavol Návrat Defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava on March 14, 2013.

Kuzar, T. Clustering on Social Web. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 5, No. 1 (2013) 34-42

We decided to focus on adoption of data clustering technique by using additional information gathered from Social Web. Partially we address topic of data visualization in user study. We analyzed approaches for processing of textual data gathered from social web by considering specific features of social web content in order to improve access to relevant information. We assume that precise content clusters will benefit users and social web data analysts as well.

In our research we studied text extraction, term selection, clustering and social web specific processing methods. We measured and evaluated the contribution of each method to the overall quality of clusters built on top of social web data. In our research we focus on pre-processing phase of web content clustering. We focus on blog articles published in Slovak language. We evaluate the impact of different data pre-processing methods on success of blog clustering. We found out that applying various text data manipulation techniques in preprocessing can improve the quality of clusters. The quality of clusters is measured by traditional clustering metrics like precision, recall and F-measure.

This paper is structured as follows: we present researches related to our study. In other part we present methods which deal with Social Web data from different perspectives. In the evaluation part we present experimental results of designed methods. Evaluation was done using our unified clustering framework and we measured impact of processing methods on quality of final clusters. In the end we discuss the possible usage of high quality content clusters in the user study.

2. Related Work

Unstructured data processing tasks are being studied from various perspectives - data retrieval, data representation, processing of massive datasets, data visualization and many others. Emergence of Web 2.0 and Social Web made the unstructured data related research even more topical.

Preprocessing. Many preprocessing concepts and algorithms - tokenization, normalization, segmentation - were researched in early machine processing era. Currently there is a need to adopt these concepts to fit the needs of social web. In web environment is text processing challenging task due to many abbreviations, misspelled words or colloquial expressions. Authors in [10] discuss trivial task as tokenizing. They explain that this task is not always simple, for example due to abbreviations. But some special tasks require more sophisticated tokenizer as in research [12], where authors analyze method for Twitter messages tokenizing which contain very high number of abbreviations and special characters. Tokenized text got normalized. Usage of normalization technique is not only language dependent but also application domain and machine learning method dependent, according to [11], [19]. Different techniques are applied which can be applied alone or in combination. Other aspects of social web processing are related massive datasets.

Feature selection. Feature selection is closely related to term extraction and many researched do not distinguish between those tasks. Feature selection algorithms were introduced and tuned in past decades to process off-line sources like newspaper, literature or legal documents. Many feature selection approaches were studied, N-grams, noun phrases and with POS (Part of Speech) tags in [9] or based on word relationships derived from corpus or linguistic resources [3]. Social Web brought additional taxonomic systems based on Wikipedia or based on social tagging - folksonomy was introduced in [16]. Social Web brought a possibility to analyze and process the documents by considering their wider context and used interactions - comments, tags, web searches, web usage and trending, hyperlinks, statuses. User comments were studied in [2, 17, 13, 20]. Other authors [19] use basic TD-IDF method. Many term selection methods are combined with TD-IDF method, some rely on statistical methods.

Statistical methods. Usually the number of extracted terms is too high and just the most important terms need to be selected. According to the literature, several statistical dimensionality reductions methods could be applied. Authors in [11] used PCA (principal component analysis) for dimensionality reduction.

Another method useful for dimension reduction is Latent Semantic Indexing (LSI) which uses mathematical technique called Singular Value Decomposition (SVD). Authors in [4] used LSI for multilingual documents clustering. Authors in [8] introduced Probabilistic Latent Semantic Indexing (pLSI) model which outperforms LSI in many aspects.

Authors in [14] introduced LDA probabilistic topic model which can be used also for dimensionality reduction. LDA uses a sampling technique in order to discover K topics in the corpus. Figure 5 describes LDA algorithm in the plate notation. LDA has many extensions and modifications as can be read in [1]: inference techniques collapsed Gibbs sampling, variational Bayesian inference collapsed variational Bayesian inference, maximum likelihood estimation, and maximum a posterior estimation. Some more can be found in [18]. According to [7] LDA was widely adopted and used in different text-related tasks as text classification, topic-specific keyword finding, topic relationship mining or relevance feedback for information retrieval.

Authors in [6] presented Latent Semantic Association (LaSA) model which can be learned from the corpus using hidden topic models like LDA or pLSI. Authors in LaSA model preprocessed the input text where they identified several words before and several words after the main word and all the surrounding words considered to be context of the main word. For all the candidates they picked surrounding features from the corpus. After that they processed the text using LDA and got the topic vectors.

Many of models mentioned above were already studied in conjunction with document clustering. Authors in [15] compared pLSI and LDA as dimensionality reduction techniques for clustering. Based on the results of their experiments, they argue that LDA and pLSI can be used for vector dimension reduction without degrading cluster quality.

We understand document representation as very important in many text related tasks. We decided to represent documents by LDA topic model. Further direction of research could aim to creation of a model of discussion behavior on blog portal in order to effectively plan article publishing and in order to track the activity regarding

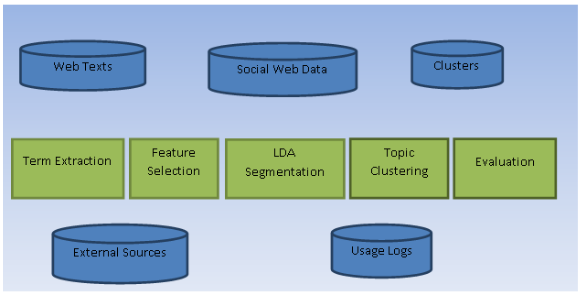


Figure 1: Building block of the processing framework.

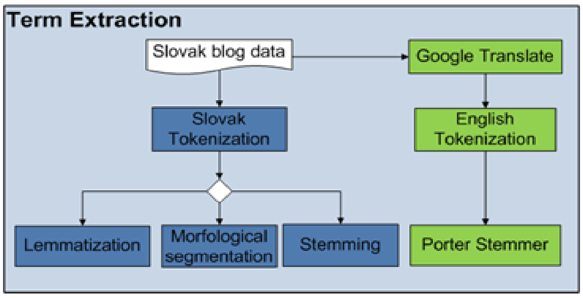


Figure 2: Different term extraction approaches.

the industry segments on the web. In order to achieve our goal we need to cluster or classify the documents.

Web related processing tasks. Previously described approaches focus on text processing. Other approaches work with information on social web on higher level of abstraction. Authors in [17] used the tag information that is associated with the web pages in order to improve clustering performance. Their method combines features from articles with tag features. Authors in [2] presented a recommendation framework which can recommend articles considering not only the original news itself but also the thread of changing comments. Authors in [5] presented blog ranking algorithm based on reading, browsing and commenting activities. Their evaluation was based on prediction of commenting activity. Their results confirmed that blog comments are valuable data source for blog ranking. Authors considered the number of comments to be one reasonable way to measure the attractiveness of a blog. Authors said that comments made by authoritative bloggers are more valuable. According to the authors blog has high commenting score if many high score bloggers make comments on it. Current researches rely on massive data storage and processing using distributed architectures.

Visualization and analysis. Gathered and processed data needs to be analyzed or visualized in order to squeeze the useful knowledge from the data. Processed data is used in various ways: recommenders, topic aggregators, trend finders. Content based recommenders are often used in case that the content changes are very fast (e.g. online news) or in environments with many components where it is hard to find similar users based on traversal patters. Research in recommenders is progressing fast but the recommended content is still not fully reflecting the needs of the users.

3. Methods

In our research we focus on processing of data gathered from social web. We gathered not only text itself but also the relations. Our framework considers different aspects of processing of data on social web - web data preprocessing with accent on Slovak language, term selection on the web and clustering which takes web data into account. We shaped our methods on subset of social web data - news articles, blog posts, user comments. Main aim of data processing is to build precise content clusters which can be analyzed and visualized further.

Processing framework handles several tasks - preprocessing, feature selection, topic segmentation and topic clustering. Within each task few methods are designed. Methods transform input data format into output data format. Basic elements of our framework are depicted on the Figure 1. Data elements are blue, processing units are green. There are 4 types of data units present during processing:

- Web documents - blog posts and news articles published in Slovak language. Web texts were manually annotated because of evaluation of the methods.
- Web document relations - information about explicit relations between blog authors and web comments related to the documents.
- External sources - Wordnet, Eurovoc, lemmatization dictionary
- Content clusters - clusters created on top of processed web document

Processing tasks and methods are described below in this section.

3.1 Term extraction

We understand term extraction as processing step which cuts input text into terms and handles the noise in the input data. In term extraction phase we focus on the morphology of the language. This task is challenging especially in less studied and morphologically rich languages like Slovak, Czech and other Slavic languages. In our research we focused on known approaches which we adjusted to Slovak. As result of term extraction we expect words in basic - morphologically simple - form. Our aim is to convert text into basic tokens which will represent the text. We studied 4 approaches to term extraction as depicted on figure below - lemmatization, segmentation, stemming and English stemming.

Lemmatization. Specific feature of Slovak are the accents. We used lemmatization dictionary and also dictionary with removed accents. With such dictionaries we lemmatized a dataset with web articles and web comments.

Stemming. Slovak language uses high number of suffixes. We created a method, which can be considered for very simple variation of morphological segmentation. Our method is based on grouping lexically similar terms into one term. We calculated lexical similarity on terms longer than three characters. If two terms are equal on more on 75% of term length, they are mapped to the same lexical term. This method is evaluated indirectly using clustering-based evaluation framework.

Morphological segmentation. We applied an algorithm for morphological segmentation on dataset where we filtered terms shorter than three characters. We randomly selected 100 terms and manually evaluated the quality of segmentation.

English translation approach. Machine translation services work are being continuously improved. We created a method where translated Slovak web articles into English, we tokenized the text and we used Porter stemmer in order to produce the stems.

We defined term extraction phase as tokenization and normalization. It needs to be considered will influence feature selection especially in case of multi-words and named entities. In specific cases term extraction and future selection should be performed together.

3.2 Feature selection

Dataset based on natural language consists of (even after precise preprocessing) very high number of units which differ in importance and there are also many synonyms and homonyms (polysemy and homonymy) among these units. The aim of the feature selection is to select a number of representative features which would represent the input documents.

In feature selection we designed three methods which decrease the number of distinct terms by considering some external or additional knowledge - web relations in form of user comments, semantic dictionaries or semantic taxonomies.

In two feature selection approaches - taxonomy based and named entiry recognition - advanced preprocessing was ommited because both methods rely on unprocessed external taxonomies and data. Web comment data set was preprocessed together with input documents.

3.2.1 Taxonomy based FS

Documents are usually represented as vectors of terms without considering relations between those terms. Term vectors can be enriched by relation by using semantic taxonomy. Taxonomy reflects semantic relations between terms, it can be created manually or inducted from big dataset. We understand the taxonomy as an external knowledge. There are several manually created taxonomies currently available, e.g. Wordnet or Eurovoc. Our method is based on grouping semantically related terms based on Eurovoc taxonomy which is available in Slovak language.

One word approach. It uses tokenized dataset. We consider whitespaces as delimiters. We look for identical terms in tokenized dataset and in taxonomy. In case that the terms match we replace the terms by the parent element from the taxonomy.

N-grams approach. This approach is more complex. We did not use tokenized dataset as described in previous approach. We generated 1-grams, 2-grams and 3-grams from the dataset. Afterwards we matched dataset n-grams with elements from the dictionary. Then we did another loop over matched pairs. In this loop we checked if the elements in the pairs are equivalent. We used the following condition - n-gram from dataset is equivalent to expression from the taxonomy when is it is possible to create a pair between another n-gram from the same article with another expression from the taxonomy on the same level.

3.2.2 Named entity recognition

. Name entities have to be process in specific way as there are many multi word expressions.

This approach focuses on selection of names entities from the text. In our research we focused on component-based approach. Our named entity recognition framework consists of 5 components: Wikipedia component, component based on Regular Expressions, online dictionary of named entities, Google component and local dictionary component.

Most of the components rely on external knowledge - Wikipedia, dictionary of named entities or Google services. We mentioned several times in this work that we focus mainly on Slovak language on social web. Some components rely on dictionaries in Slovak language - dictionary with Slovak names and online dictionary. Google and Wikipedia are general well know services which work quite well also with Slovak language.

3.2.3 FS based on Web comments

Our research focuses primarily on user-generated textual data gathered from Social Web. Web comments are valuable source of such user-generated content. Aim of our method is to extend web documents with information from the web comments. Web comments data is usually noisy - comments with poor grammar or spamm comments. We designed a method which select only high quality comments from the dataset. Method for selecting high quality comments considers quality of the author, quality of the comment as relation to other comments and quality of the comment itself based on text analysis.

UserRank is based on PageRank algorithm and user behavior characteritics. User who has many replies on his comments we consider as more important.

$$U(i) = \frac{(1-d)}{N} + d * \sum_i W_u(j,i) * U(j) \tag{1}$$

CommentRank is based on the structure of the discussion. Comments with many replies have higher impact than the comments with only few replies.

$$C(i) = \frac{(1-d)}{N} + d * \sum_i \frac{C(j)}{N(j)} \tag{2}$$

FeatureRank is based on the comment itself. Frequency of usage of questionmarks, exclamations, stars, capital letters, web links a emoticons was considered. Sum of frequencies was divided by length of the comment not to compromise longer comments. Higher FeatureRank means that comment contains many special characters and such comments are usually of low quality.

$$F(i) = \frac{f_o + f_v + f_h + f_l + f_s}{n_w} \tag{3}$$

TotalRank is calculated as sum of all the characteristics. Comments from influential users and with many replies are increasing the rank of the comment.

$$UserRank + CommentRank - FeatureRank \tag{4}$$

Comment selection approach is based on filtering high quality comments from the dataset. Comments were ordered based on TotalRank and only 20% of top comments were used for further processing.

3.3 Document - Topic segmentation

Previous methods rely on some kinds of dictionaries or on social web knowledge in order to pick most representative features. In related work section we presented several approaches to feature dimensionality reduction. Other approaches to select representative features are based on statistical methods. Dimensionality reduction is a commonly used step in machine learning, especially when dealing with a high dimensional space of features [Feature selection for dimensionality reduction]. We included into our methods also statistical methods which are able to operate over text dataset but they are not constrained to operate over text dataset. We use LDA in order to transform feature vectors into topics. LDA methods consider feature distribution or co-occurrence of features and significantly decrease the document feature vectors. We used LDA in order to decrease the dimensionality of the features.

$$Document(f_1, f_2, ..., f_n) \rightarrow Document(t_1, t_2, ..., t_{20}) \quad (5)$$

Most important output LDA algorithm is document-topic matrix. Number of topics was experimentally set to 20. We considered this number as optimal for our dataset. After running LDA over dataset we got article-topic matrix which we used as input for clustering algorithm.

3.4 Topic clustering

LDA method used for dimensionality reduction transformed documents represented by features into documents represented by topics. We used topics instead of features in clustering algorithm because number of distinct would be too high. On the other hand, number of topics was set to 20 what was a solid base for applying the clustering algorithm.

$$Document(t_1, t_2, ..., t_{20}) \rightarrow C_i \quad (6)$$

As a result of clustering method we got the input documents organized in defined number of clusters. Number of clusters was defined according to the dataset characteristics.

3.5 Postprocessing

As we already described, web documents contain not only content itself but also user interactions are part of it. We used this data in order to create more precise clusters in phase called postprocessing. Postprocessing method employs articles, comments, users or clusters and the connections between them into process of building high quality content clusters. We are proposing a method which creates clusters by considering article - commenter matrix and a method which combines traditional content based clustering with article - commenter matrix.

Clustering based on comments. Clustering method based on implicit relations creates clusters of the blogs according to commentators who commented the blogs. This method is based on our assumption that similar commentators comment similar blogs and we also assume that blogs have characteristic themes and characteristic commentators. This method does not rely on language processing.

Our preliminary explorations are depicted on Figure 2 where we represented blogs and commentators as undirected graph G(V, E), where V=$\{v_1, v_2, ..., v_n\}$ stands for blogs and commentators and E=$\{e_1, e_2, ..., e_n\}$ stands for edges. We created weighted and unweighted variations of blogcommentator edges. In unweighted case there is an edge between blog and commentator when the commentator posts at least one comment to the blog. In weighted case we counted the number of comments which the commentator posted to the blog. According to the visualization depicted on Figure 2 it is obvious that blog-commentator edges create clusters.

In order to provide precise evaluation of the social clusters we represented each blog B_i from blogs B as vector of its commentators C = $\{c_1, c_2, ..., c_{|c|}\}$. In weighted case N_{ti} is the number of times the commentator ct posts a comment related to the blog bi. In unweighted case N_{ti} is 1 when commentator c_t appears at least ones in the discussion related to the blog bi and 0 in other cases. Then were the blog vectors B_i clustered using K-means algorithm into social clusters S.

Combined method. Previous method relies exclusively on implicit relations between commented blogs and commentators. Our third method uses implicit relations between commentators and blogs in addition to the content clustering method. It combines content clustering approach with clustering based on implicit relations in comments. Algorithm which combines both approaches consists of six steps:

1. *Loop.* Repeat content clustering method 10 times on whole dataset and create the content clusters G.
2. *Select.* Identify 10% of blogs in dataset which are most likely changing the clusters and put them into set U.
3. *Cluster.* Cluster blogs which are represented by their commentators and create clusters S.
4. *Compute.* For each item uk from U, find social cluster S_1 where u_k is from S_1 and find G_k from G with minimal HammingDistance(s_1, g_k) and assign item $u_k \rightarrow G_k$. HammingDistance function counts the number of positions where are corresponding values different.
5. *Update.* Update content clusters for each item u_k from U.
6. *Evaluate.* Evaluate updated clusters from G by comparing with manually annotated dataset.

In our research we studied statistical methods for feature selection and we found out that they are not precise on small dataset. Postprocessing method overcomes drawbacks of statistical methods by considering comments related to the web articles and blog posts.

4. Findings and evaluation

Our research focuses on clustering of web documents by considering knowledge hidden on social web and some external data as well. We designed and evaluated methods which operate in various stages of content processing (term extraction, feature selection, postprocessing) as described in previous section.

4.1 Evaluation approach and dataset description

Methods are evaluated against manually annotated dataset. Our clusters represent documents and articles with similar

Table 1: Dataset characteristisc

Nr.	Characteristics name	Value
1	Number of commentators	34336
2	Number of comments	2733278
3	Number of blogs	27999
4	Average number of comment replies	1,38
5	Number of direct comments	715058
6	Average number of replies on blog	715058
7	Max number of blog comments	1858

content. Some methods were evaluated separately; some were evaluation as part of clustering framework. Experiments were evaluated using Precision, Recall, F-measure which are defined as follows:

$$Precision = \frac{truepositives}{truepositives + falsepositives} \quad (7)$$

$$Recall = \frac{truepositives}{truepositives + falsenegatives} \quad (8)$$

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \quad (9)$$

Dataset description. In our experiments we work with two kinds of data - blogs and comments. From our blog dataset we manually picked 200 representative blogs and labeled them manually using 4 labels. Characteristics of our comments dataset are provided in Table 1.

Described dataset was used for evaluation within entire processing framework. Some specific components were evaluated separately on specific datasets.

4.2 Experimental results

Aim of this chapter is to present the results of most important experiments performed using proposed social web processing framework. Experiments show potential and drawbacks of designed processing framework.

4.2.1 Preprocessing evaluation and taxonomy usage

We proposed several preprocessing methods. In experimental evaluation we measured the impact of different preprocessing methods on quality of content clusters. We performed experiments on two sets: on 4 topics and on 10 topics. As depicted in the Figure 7, applying lemmatization processing F-measure increased in both sets. Usage of Eurovoc and lexical classes was successful just in case of 4 topics and it decreased the F-measure in case of 10 topics. The reason of decrease of F-measure in case of 10 topic dataset is the domain specific dictionary. And the dictionary does not cover all of the 10 topics. Lexical stemming decreases number of distinct forms. But it can increase ambiguity of the processing and degrade quality of content clusters as it was measured in case of 10 topic dataset.

We proposed to include external knowledge into document processing. We found out that domain in both cases need to be considered precisely. There is a challenge to find domain specific dictionary in case of rarely used languages as Slovak.

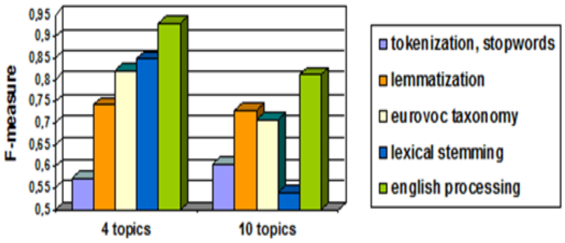


Figure 3: Different term extraction approaches.

English processing. In other setup of experiment we focused on English based processing. Clustering of articles in our dataset was successful in case of 4 topics and in case of 10 topics as well. Main reason of success is that Slovak to English translation works quite well and that English is morphologically not as rich as Slovak. We could apply Porter stemmer and decrease the number of word forms. At the Figure 7, English based processing is represented by the very right bar. English processing significantly overcome processing in Slovak language. It is obvious that proposed preprocessing methods and rather simplistic and there is much space for improvements and fine tuning.

We found out that preprocessing task considerably influence the quality of the content clusters. This influence is obvious in caseased on the experiments we of processing in English using Porter stemmer. The gap between processing in Slovak and English is significant, so we suggest fine tune Slovak preprocessing methods in order to build high quality content clusters.

4.2.2 Feature selection based on web comments

In this section we present results of experiments done on comment based feature selection method which was described in [3.3.3]. Experiment was done on subset of our dataset with 6 manually annotated categories. In order to evaluate proposed method we designed three experiments:

1. Selected articles were processed by basic preprocessing, LDA reduction and topic clusters were created.
2. Top comments related to the articles were selected using proposed method. Only comment features were processed further
3. Articles and related top comments were used for further processing.

Experimental results are depicted on the figure 4. Setup where only comments were processed had the worst results. Setup where articles were processed together with related comments slightly overperformed initial scenario where only articles were processed.

According the experimental results, clustering of comment features themselves brings worse results then clustering of the articles. The reason is that in some cases the content of the discussion in comments does not correspond with the content of the article. But in most cases the comments extend the basic idea of the article. That is why overall quality of clusters can be improved by combining both approaches. In comment-based feature selection we found out that comments feature vector contains very

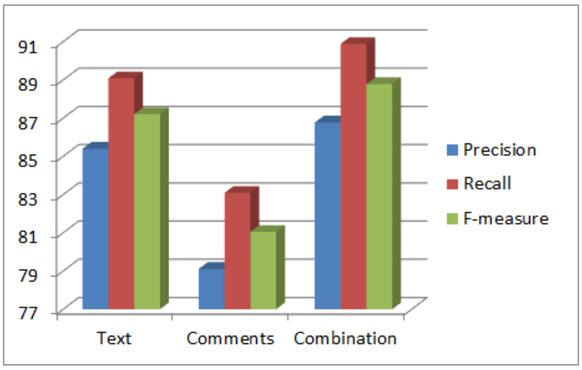


Figure 4: Different term extraction approaches.

Table 2: Comment based clustering results

Nr.	Method name	F-measure
1	Weighted edge based	0,42
2	Unweighted edge based	0,52

frequent features which are not present in article feature vector. That could help to give broader insight to the articles.

4.2.3 Postprocessing evaluation

We evaluated both proposed postprocessing methods: clustering based on commnents and method with combined content and comment clustering. Results of the first experiment are presented in the table 2. It may be surprising the unweighted experiment setup overperformed weighted. It is so because comment based clusters differ from content based clusters. And by weighteing the relations the differences between both comment based clusters and annotated dateset were more significant.

In other setup where content clusters and comment clusters were combined, we run repeated the experiment 10 times and calculated F-measure after each observation. The results are depicted on the figure below. The quality of the clusters slightly increased in F-measure in most cases.

Based on experimental results our comment-based enhancement has very small but positive impact on quality of clusters. Impact of proposed method is small due to the fact that it was applied on small subset of unstable posts. Our experiment shows that a method which takes implicit ties between commentators into account benefit the clustering method but just in case that it is user as a supplement to content-based clustering. In postprocessing we focused on increasing the quality of content clusters by applying some knowledge acquired by mining the comments which are related to clustered documents. We found out that clusters purely based on implicit social connections between commenting users differ from clusters based on content. But we found out that content clusters can be improved by considering social ties in case of articles without distinct topic. Combination of content clusters with comment based clusters for articles without distinct topic improved overall quality of clustering.

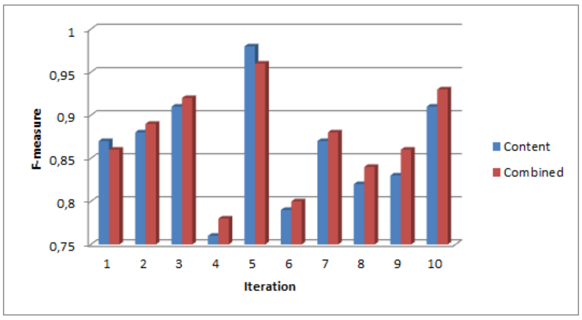


Figure 5: Different term extraction approaches.

4.2.4 Intepretation of the experimental part

Web concentrates great amount of user generated data like bookmarks, tags or comments which can be used in various text-related tasks. We designed a processing framework which aims to build high quality content clusters on top of social web data. We designed content clustering framework as set of processing tasks which influence each other and more or less impacts the quality of the content clusters. In our experiments we measured the impact of different processing methods on quality of the content clusters.

We found out that the lemmatization task significantly increases the quality of clusters. We found out that the quality of clusters can be increased also by features selection based on taxonomy. We tested our taxonomy based feature selection method on several datasets and we found out that it improved the quality of clustering just in case when our taxonomy matched the used dataset. We also found out that clustering based on translation of Slovak text into English and applying English processing methods significantly outperformed processing in Slovak. We found out that there is much space for improving Slovak text processing methods.

We evaluated usage of comments in feature selection taks and also in postprocessing task. We both cases we observed that comments comments should be processed together with the articles not separately.

We designed also named entity recognition method. This method was evaluated separately not as part of content clustering framework. We found out that method which consists of different processing components increases the overall quality of the recognition. And we found out that the overall quality of the system of components is higher than the quality of any separate component. This supports the finding learned from comment processing: increased amount of data and increased number of processing tasks has positive influence on quality of the content clusters. Main constraint in case of very complex data processing frameworks is the performnace.

Based on results of provided experiments we conclude that combination of term extraction, feature selection, clustering and postprocessing tasks lead to better clustering results. But liminations like complexity of Slovak language, lack of well defined taxonomies or scalability constraints related to processing of complex Social Web datasets should be considered.

5. The user study

Motivation of our research was to simplify information access on the Social Web. In the user study we demonstrate how content clusters help the users to get the rele-

Table 3: Click through rate based on content

Name	Clicks	Views	Rate	Type
E 145rMY	6281	97308	6.45	Specific
E 141Pae	2329	37387	6,23	Specific
E 143Swv	1705	30663	5,56	Specific
E 146ARD	1522	25485	5,97	Specific
E 147QCz	1249	19323	6,46	Specific
E 142JoY	845	16294	5,19	Specific
E 148vNA	929	15706	5,91	Specific
E 149CXm	480	9808	4,89	Specific
What's up 298vJN	800	30152	2,65	General

vant information. User study was based on analysis done on three web applications displaying the same content via there different web applications - Noticeboard, Widgetizer and Visualizer.

Noticeboard. We designed a web application where the user put notices about coming cultural events together with short descriptions and classifications. The users insert information into the noticeboard because they want their submission reach the broader audience.

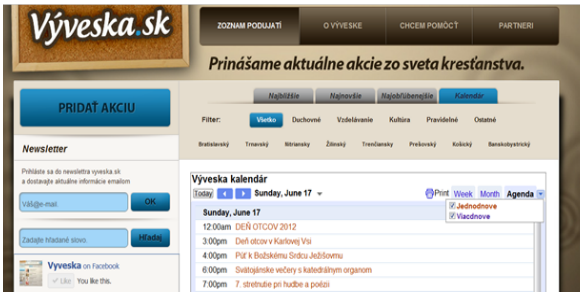
Widgetizer. Its name is derived from its main functionality - creating widgets. Widgetizer is a tool which allows the users to create HTML widgets which visualize the information gathered from RSS sources. Content of the widget is filtered based on manually specified selections. Widgetizer is able to handle techniques like personalized recommendations and targeted context adds.

We designed an information propagation architecture based on web services and syndication formats. Aim of the architecture is to distribute relevant information to the small web pages without much work and programming knowledge. Architecture has three main components - feed aggregator, datastore and data distribution channel. Information distributed through distribution channel can be filtered manually and automatically depending on source content and the content of the target web page.

According the user study we assume that the content within the same content cluster or category is more interesting for the end user as different kind of content. Table above compares user interests for general content and filtered content. User are most than twice more interested into filtered content that to general content.

Visualizer. Business is able to analyze the clusters and the activity in these clusters, business can collect moods, trends and feedback from the social web. Created segments can be analyzed using click stream analysis software (e.g. Google Analytics). Additionally, on the back end, editors are benefiting by being able to share, associate and package news (products, information) across properties, significantly increasing opportunities to monetize content. This is out of scope of this paper to discuss the monetization of the content.

Based on the evaluation, it is obvious that information consumers need the information to be organized in specific blocks. Usually users are interested in the information within specific topic and thus information organized in content clusters will help the users to orientate in social web space - no matter whether the end user just searches the web or business analyst analyses the activity and trends within a cluster.



Prihláste sa do newslettra vyveska.sk a dostavajte aktuálne informácie emailom

You like this.

Výveska kalendár

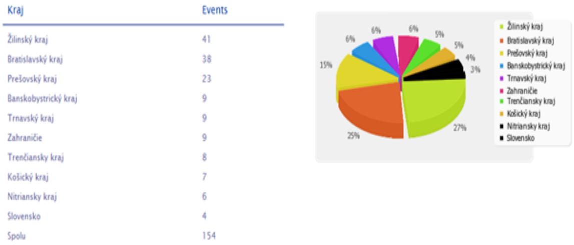
Sunday, June 17

☑ Jednodnove
☑ Viacdnove

Sunday, June 17

12:00am DEŇ OTCOV 2012
3:00pm Deň otcov v Karlovej Vsi
4:00pm Púť k Božskému Srdcu Ježišovmu
6:00pm Svätojánske večery s katedrálnym organom
7:00pm 7. stretnutie pri hudbe a poézii

Figure 6: Another approach to visualization of clusters is the calendar view.



Kraj	Events
Žilinský kraj	41
Bratislavský kraj	38
Prešovský kraj	23
Banskobystrický kraj	9
Trnavský kraj	9
Zahraničie	9
Trenčiansky kraj	8
Košický kraj	7
Nitriansky kraj	6
Slovensko	4
Spolu	154

Figure 7: Activity on clusters can be analysed.

6. Conclusions and contributions

In this paper we discussed topics related to clustering on Social Web: definition of information producer and information consumer, design of content processing framework with focus on clustering and evaluation of its components, user study which demonstrates the benefits of the content clusters for the web users.

In the introduction part of this paper we described an idea of connecting information producer with information consumer in effective way. With Web 2.0 every web user got a possibility to be a web publisher. Now web user acts usually in two roles in the same time - content producer and content consumer. User as content producer is fladding web space with enormous amount of data as statuses, comments or tags. And the user in role of content consumer tries pick up most relevant content for him. We concluded that user in role of content consumer should get the information in effective way by relying on automated data processing. Based on our research we state that social web user usually acts as information producer and information consumer in the same time. We identified as very important to receive web data in effective and thus we focused on organizing documents into clusters.

We proposed to organize social web data into content clusters can help the information consumers effectively receive the information. Main part of this paper focuses on content processing framework for building high quality content clusters. Processing framework consists of several components which operate in different processing steps. Intensive processing of unstructured data is a prerequisite of high quality outputs. We found out that information preprocessing significantly influences the quality of content clusters. Then we found out that quality of content clusters can be influenced by dictionaries but the domain of the dictionary needs to match domain of the input dataset.

We intensively analyzed web comments. We used content of web comments as extension of web documents and also we used web comments as indicator of relations between social web documents. Enriching web documents with comments had positive impact on overall quality of clustering. But we found that processing comments increases quality of clusters only in cases that comments are processed together with articles.

Then we designed several social web applications which are intended to operate on top of the created content clusters - noticeboards, widgetizer, visualizer or calendar. We assume that such application will bring new possibilities for more effective usage of social web. Based on the user study we think that business areas like Customer Relationship Management, Product Management and Marketing can benefit from knowledge retrieved from social web. The better these business areas know their target audience, the more tailor-made solutions can they prepare. In order to get internet-based information about their customers and products they need to gather, process and analyze information from the social web and integrate it into their internal enterprise applications.

Proposed methods were deployed on traditional architectures. Traditional approaches are short on scalability. This challenge can be effectively solved by big data approach which we find very promissing. As a future work we recommend to analyze social web content with considering relations on social web. And also we recommend focus on processing of very big datasets gathered from social web.

Acknowledgements. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09, the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11 (2011-2014) and the Slovak Research and Development Agency under the contract No. APVV-0208- 10. This contribution/publication is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

[1] A. Asuncion. Auai. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence.*. Press Arlington, Virginia, United States, 2009.

[2] L. B. Enhancing clustering blog documents by utilizing author/reader comments. In *Proceedings of the 45th annual southeast regional conference*, pp 94 – 99. ACM, NY, USA, 2007.

[3] G. Cao, J.-Y. Nie, and J. Bai. Integrating word relationships into language models. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp 298–305. ACM, USA, 2005.

[4] W. Ch. A latent semantic indexing-based approach to multilingual document clustering. In *Decision Support Systems*, pp 606–620. ACM Press, 2008.

[5] E. Elgersma. Personal vs non-personal blogs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp 723–724. ACM, 2008.

[6] H. Guo. Product feature categorization with multilevel latent semantic association. In *Proceeding of the 18th ACM conference on Information and knowledge management*, ACM, NY, USA, 2009.

[7] V. Ha-Thuc. Topic models and a revisit of text-related applications. In *Proceeding of the 2nd PhD workshop on Information and knowledge management*, ACM, USA, 2008.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.*, New York : ACM, 1999.

[9] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp 216–223. Association for Computational Linguistics, Stroudsburg, PA, USA, 2003.

[10] R. M. KAPLAN. A method for text tokenizing. In *Inquiries into Words, Constraints and Contexts*, pp 55–64. CSLI Publications, Standford University, 2005.

[11] T. Korenius. Stemming and lemmatization in the clustering of finnish text documents. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp 625–633. ACM Press, 2004.

[12] G. Laboreiro, L. Sarmento, J. Teixeira, and E. Oliveira. Tokenizing micro-blogging messages using a text classification approach. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pp 81–88. ACM, New York, NY, USA, 2010.

[13] Q. Li. User comments for news recommendation in forum-based social media. In *International Journal of Information Sciences, Volume 180 Issue 24*, pp 4929–4939. Elsevier Science Inc. New York, NY, USA, December 2010.

[14] B. M. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(0), 2003.

[15] T. Masada. Comparing lda with plsi as a dimensionality reduction method in document clustering. In *Proceedings of the 3rd international conference on Large-scale knowledge resources: construction and application*, pp 13–26. Springer-Verlag Berlin, Heidelberg, 2008.

[16] A. Mathes. Folksonomies - cooperative classication and communication through shared metadata. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp 0–0. Computer Mediated Communication - LIS590CMC, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December 2004.

[17] C. S. Ranking weblogs by analyzing reading and commenting activities. In *WI-IAT '09: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, pp 442–449. IEEE Computer Society Washington, DC, USA, 2009.

[18] Z.-Y. Shen. Collective latent dirichlet allocationdata mining. In *ICDM '08. Eighth IEEE International Conference on.*, IEEE, 2008.

[19] L. Sun. User-driven development of text mining resources for cancer risk assessment. In *Proceedings of the Workshop on BioNLP*, pp 108–116. ACM Press, 2009.

[20] J. Zhang. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th international conference on World Wide*, pp 221 – 230. ACM New York, NY, USA, 2007.

Selected Papers by the Author

T. Kuzár, P. Návrat. Burst Moment Estimation for Information Propagation. In *Advances in Intelligent Web Mastering - 2: Proceedings of the 6th Atlantic Web Inteligence Conference - AWIC 2009*, pp 147–154, Czech Republic, 2010. Springer

T. Kuzár, P. Návrat. Preprocessing of Slovak Blog Articles for Clustering. In *Proceedings of the 2010 IEEE/WIC/ACM Int. Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2010) Workshops Proceedings*, pp 314–317, Canada, 2010. IEEE CS.

T. Kuzár, P. Návrat. Slovak Blog Clustering Enhanced by Mining the Web Comments. In *Proceedings of the 2011 IEEE/WIC/ACM Int. Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03 (WI-IAT 2011)*, pp 293–296, France, 2011. IEEE CS