Mining of relevant and informative posts from text forums

Kseniya Buraya, Vladislav Grozin, Vladislav Trofimov, Pavel Vinogradov, and Natalia Gusarova

ITMO University

Abstract. In the modern world, the competitive advantage for every person is the possibility to obtain the information in a fast and comfortable way. Web forums occupy a significant place among the sources of information. It is a good place to gain professionally significant knowledge on different topics. However, sometimes it is not easy to identify the places on the forum, which contains useful information corresponding user demands. In this paper we consider the problem of automatic forum text summarization and describe the methods, which can help to solve it. We study the difference between relevance-oriented and useful-oriented query types. We will describe our dataset, that contains over 4000 of marked posts from web forums about various subject domains. The posts were marked by experts, by estimating them on a scale from 0 to 5 for selected query types. The results of our study can provide background for creation informational retrieval applications that will decrease the time of user's searching and increase the quality of search results.

1 Introduction

The number of various informational resources is constantly growing nowadays. Upgrading knowledge in particular areas often becomes very time-consuming and difficult. Also, it is not so trivial to obtain basic knowledge in some new subject for the person who is not very competent in it. Thus, it is very important to have quick and comfortable access to information. First of all, it is increases the possibility to obtain knowledge about the most important aspects of the area of interest. Web forums are among the most important resources for the acquisition of professionally significant information. There people communicate with each other by creating the threads, that are dedicated to a specific topic, and lead the discussion by writing posts to them. Thus, a thread presents a well-formed user-discussion process on the declared subject.

As a resource of professionally significant information, compared to traditional educational resources and scientific publications, forum has the following advantages:

– the forum contains the most up-to-date information on a topic. Specific technological solutions are often formed in user's discussions, while the publication of the same information requires a long time;

- forum posts represent the experience of people who are directly using the specific technologies and have both positive and negative experiences. Such information is practically not available in the official documents;
- the information on a web forum is presented in a structured manner. It extends the capabilities of the informational search;
- the way of presenting information on a forum has more freedom in describing details, contains emotional evaluations and different types of visualization;
- the information on a forum reflects the collective opinion of the professional community.

At the same time, there are some disadvantages of using a web forum, as a source of professionally significant information:

- information redundancy — a large amount of repetitive, highly emotional and professionally irrelevant information;
- topics drift — changing the originally declared theme to others;
- the disadvantages of language — incomplete sentences, the differences in the understanding meaning of concepts in separate posts. This makes it difficult to analyze forums in foreign languages.

Consider the typical situation. Someone wants to learn about technology, which can be useful in his/her recent activities. Search query leads him/her to a forum, where the technology is discussed. The questions are: is there enough information on the forum for the detailed acquaintan with the area of interests?; Which posts contain really professionally significant information? It would be helpful to obtain the answers to these questions and then to study selected posts in details. This means that we have the problem of the automatic offline summarization of the most important posts, which contains professionally significant information. This task also becomes more meaningful when a forum is in unknown foreign language and available only through translation.

Authors of [3] proposed different approaches for text forums summarization. The most powerful methods for this task are the machine learning methods [5,25]. However, there is a high number of different methods in machine learning, and the selection of the most efficient ones is a problem. One of the main things in text and forum summarization is the extraction of keywords [2]. Keywords extraction methods are divided into two categories: selection of words from a predefined vocabulary or taxonomy by the document content, and extraction of keywords directly from the documents in analysis [16]. The methods of the second group are also divided into several categories [4,32]: machine learning methods, linguistic methods, graph methods, statistical and heuristic methods. Statistical methods are based on the computation of different statistics of documents, including the frequency of word occurrences, tf-idf, n-grams and so on [25]. Heuristic methods [28] allow developing the structure of the document by using characteristics such as position of the word in the document, existence of formatting of the elements, document fragment length, etc. According to the article [3], the main tasks in forum summarization are sentiment-analysis, allocation of facts from the documents, analysis of user activity. However, at the

same time, the problem of highlighting of professionally significant information is not presented even in its formulation.

Thus, in this paper, we consider the problem of automatic summarization of web forums. Our goal is to study the methods of selection forum posts, which contain professionally significant information.

2 Related works

There are different approaches to the problem of text summarization. They can be divided into extraction-based and abstraction-based [26] summarization. Also, there are single-document and multi-document approaches. The majority of works in the area of forum summarization use extraction-based techniques and single-document approach [22]. Extractive forum summarization tasks are divided into generic summarization (obtaining a generic summary or abstract of the whole thread) and relevant query summarization, sometimes called query-based summarization, which summarizes posts specific to a query [12].

We found several types of research close to our work in literature. Authors of [11] studied reviews posted on the web assessing "Review Pertinence" as the correlation between review and its article. Authors of [29] considered the sentence relevance and redundancy within the summarized text. Their maximum coverage and minimum redundant (MCMR), text summarization system, computed sentence relevance as its similarity to the document set vector. This idea was also used in [30] for cross-lingual multi-document summarization.

Some articles [30,21] were devoted to comparing system effectiveness and user utility. Authors of [20] compared traditional TREC procedure of batch evaluation and user searching on the same subject. Authors of [21] confirmed that test collections and their associated evaluation measures did predict user preferences across multiple information retrieval systems. They found that NDCG metric modeled user preferences most effectively.

To sum up, there are no articles with the in-deep study of the problem discussed in our article.

3 Methods

On the one hand, it is proposed to consider certain aspects of user's informational needs. The author of [24] has defined six possible assessment levels for information systems, where the first three were referred to measuring system performance (such as speed of the processing the query, matching the query and document content), the last three levels corresponded to user-oriented evaluation (including feedback, context, social and cognitive matching of query and document and etc.). The author of [14] uses the following measurements: (1) user's characteristics (gender, age, etc.); (2) the interactive parameters (the number of sent requests, the number of viewed documents, etc.); (3) the quantitative characteristics of query results (accuracy, completeness, NDCG, etc.); (4) the qualitative users characteristics (declared by experts).

On the other hand, there are some international projects [25,9,18,13] for evaluation of information retrieval systems, based on the user's information needs. Each project contains an annotated collection of documents (mainly in the style of news), divided into groups of informational needs (tracks). The results of system performance are evaluated by laboratory experts following strong established and context-limited queries statements that limit the applicability of this approach to practical problems.

The analysis of this approaches shown, that the lower levels of classification that described by Saracevic [24] and Kelly [14] can be evaluated by traditional informational retrieval systems quality metrics (such as F-measure, NDCG, etc.). However, the possibility of using them for upper levels associated with the formulation of appropriate information request proposed to expert. In this case, to evaluate the efficiency of extraction of professionally significant information, we formulate information needs in the form of problem-oriented queries and use different contexts for their evaluation. On the one hand, this approach is consistent with the structure of scaling requests adopted in TREC [9]. On the other hand, their contents cover the real user's informational needs when searching for professionally significant information. So our target variables will be Informativeness and Relevance. Formal criteria for marking them up are listed in Table 1. It is obvious that binary evaluation of the quality of extraction of professionally significant information would be too coarse-grained. For expert marks of informativeness and Relevance, we use the six-level scale, constructed in a similar way, that described by Elbedweihy [9]. This allows us to consider the measured values as categorical or continuous in the interval [0, 5], depending on selected problem formulation: classification or regression. Also, our experts were given explicitly formalized instructions on how to mark up posts, using a strict and formal scale from Table 1. In order to avoid subjectivity and bias, we've involved several experts.

3.1 Quality estimation

Widely used metrics such as F-score, recall/precision, and others are not applicable in our context. Although these measures are commonly used in both IR and semantic search evaluations, their main limitation is that they must be used with a binary scale. Because in our work we are using a non-binary scale, it makes more sense to follow the recommendations of the Elbedweihy [9] and use cumulative gain metrics to evaluate retrieval system quality. We used normalized cumulative gain, that is quality metrics, based on a comparison of the calculated position of the post with its position in the perfect sorting by expert marks. It's calculated using formula:

$$NDCG_N = \frac{DCG_N}{IDCG_N},$$

where

$$DCG_N = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{\log_2(i)},$$

Table 1: Formal markup criteria

Parameter	Context	Value	Comment
Informativeness	Display posts that contains objective, interesting and professionally significant information on request	0	Post contains no useful information
		1	Post gives some useful information, but most of it is not useful
		2	Post gives a little amount of useful information
		3	Post contains useful information, but explanations and arguments are missing
		4	Post contains useful information, but explanations and arguments are incomplete
		5	Post contains a lot of useful information with rich explanations and arguments
Relevance	Display posts that contains semantically close information om request	0	Post is completely irrelevant to the query/topic
		1	Posts theme weakly intersects with query/topic
		2	Post contains mostly irrelevant information, but some parts of it are relevant
		3	Post contains mostly relevant information, but some parts of it are irrelevant
		4	Post is relevant to the query/topic, but contains some extending information
		5	Post is completely relevant to the query/topic

N is the size of resulting set (how many documents to retrieve), rel_i is true value of target variable (relevance or informativeness) of i-th post in the retrieved set, and $IDCG_N$ is the maximum possible value of DCG_N for specified forum and for given N, i.e. DCG_N for an ideal algorithm.

To ensure model stability, we used bootstrap-like method. The data was resampled with replacement, then it was split into test and train sets. After that, models were fit, and model qualities were estimated. This process was repeated 200 times, and model qualities was averaged and confidence interval was calculated:

$$StD_{NDCG} = \sqrt{\frac{\sum_{i=1}^{k}(NDCG_i - \overline{NDCG})^2}{k}},$$

where k is the number of bootstrap splits, $\overline{NDCG}$ is the average of $NDCG$ values for k bootstrap steps, $NDCG_i$ that is the $NDCG$ value on the kth iteration of bootstrap.

3.2 Features

There are various methods for feature extraction proposed in the literature [6,27,19,23]. Based on our previous work, we have made problematic-based feature selection (Table 2). These features are divided into four groups: (1) the position of the author of the post among other users (his position in the social graph); (2) the position of the post in the thread; (3) text features; (4) the emotional evaluation of the post. In our study, we used expert marks for evaluating the emotional component of posts. We calculated the features from the first group in two ways to determine the possible relations between emotional evaluation and the values of the target variables for each post using weighted (sentiment graph) and unweighted (non-sentiment graph) graphs.

Table 2: Features

Type	Feature and it meaning
Post author graph features	Betweenness, non-sentiment graph (Author's social importance) inDegree, non-sentiment graph (How many times author was quoted) outDegree , non-sentiment graph (How many times author quoted someone) Betweenness, sentiment graph (Author's social importance) inDegree, sentiment graph (With which sentiment author was quoted) outDegree, sentiment graph (Author's quotes sentiment)
Post author features	Number of threads author is participating in (Author activity)
Thread-based post features	Position in thread (Chance of off-topic) Times quoted (Post impact on forum)
Text features	Length (Number of arguments and length of explanations) Links (Number of external sources/images) Sentiment value, calculated using sentiment keywords (The emotional evaluation of post) Number of query keywords (Topic conformity) Most used topic keyword count (Topic conformity)

3.3 Models and parameters

There are various machine learning methods and their algorithmic implementations nowadays. In the academic literature, there are different principles of their classification. Also, there are constantly updated ratings, which are made by users and developers. As it was shown above, the result of extracting professionally significant information is determined by the context of the request and the type of evaluation metric, which is associated with the formulation of machine learning problem. So we used two classifying attributes for selection

of methods and models of machine learning problem: the type of ML problem (classification/regression) and the target variable (informativeness/relevance).

The selected machine learning algorithms and their paramters are listed in Table 3. For each algorithm we selected its regression and classification modes.

Table 3: Models parameters

Model	Algorithm	
	Classification	Regression
Regression	Logistic	Linear
	By default	By default
Support Vector Machine	LibSVM	LibSVM
	By default	By default
Decision Tree	J48	M5P
	Use reduced error pruning: true	Build regression tree/rule rather than a model tree/rule: true
K-nearest neighbors algorithm	IBk	IBk
	Neighbors number: 5; Weight neighbors: by the inverse of their distance; Neighbour's number selection: hold-one-out evaluation; Minimization parameter: mean squared error	Neighbors number: 5; Weight neighbors: by the inverse of their distance; Neighbour's number selection: hold-one-out evaluation; Minimization parameter: mean squared error
Neural Network	MultilayerPerceptron	MultilayerPerceptron
	Learning Rate for the backpropagation algorithm: 0.001; Momentum Rate for the backpropagation algorithm: 0.001; Number of epochs to train through: 5000; Percentage size of validation set: 20	Learning Rate for the backpropagation algorithm: 0.001; Momentum Rate for the backpropagation algorithm: 0.001; Number of epochs to train through: 5000; Percentage size of validation set: 20
Naive Bayes / Gaussian model	NaiveBayes	GaussianProcesses
	Use kernel density estimator: true	By default

To analyze the impact of the query context on the quality of using machine learning methods we use the following models:

– Multiple Linear Regression (LM). Attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. In connection with a sufficient amount of data and weak correlation between features, we use non-regularized model [8].

- Stochastic Gradient Boosting (GBM) which is a model capable of capturing nonlinear dependencies. We used three CV folds to estimate the best amount of trees; a number of trees were capped to 2000, and shrinkage factor was 0.001. Indirection level value (number of splits for each tree) was set to 3 [10].
- Latent Dirichlet Allocation (LDA) that is robust interpretable model splits available posts into subsets (topics) according to their texts using bag-of-words approach. Each topic can be interpreted as a set of keywords, and we used the presence of these keywords to estimate target variables. Comparing the keywords sets in formed topics and texts of post we can distinguish the posts, which are corresponding to the specific query context [7].
- Cumulative link model (CLM). Also known as ordered logit model. This is modified ordinal version of multilogistic regression that makes use of the fact that we have several ordered classes [1].
- Word2Vec. This is a parametric model that are used to produce word embeddings. It assigns high-dimensional vector to each word in such way that words with similar meaning have similar vectors. In our experiments we use complete Russian National Corpus[1] model for Russian language and Google News Corpus for English[2]. For each forum post, we do the following steps: split post and query into lexemes, and calculate semantic similarity between each lexeme and user query. After that we rank each post by sum of these similarities [17].

For more detailed analysis we also compare different methods of keywords extraction:

- Most Used Keywords. The model considers posts as a big set of words and selects the most frequent ones.
- Hclust. The model considers thread text as "Document-Term"matrix and forms a hierarchical classification of words. Clustering is the process of partitioning a set of objects into subgroups (clusters) according to proximity or some other criteria. Formally the problem is posed as follows: let $X = \{x_1, x_2,...,x_n\}$ be a finite set of objects; Y is a set of clusters. Then $\rho(x, x')$ is the distance function between objects x and x'. We are to partition the sample X into disjoint subsets (clusters) in such a way that each cluster consists of objects that are close according to the metric , and object of different clusters are substantially different in this metric. Each object $x_i \in X_n$ is assigned with the corresponding cluster index y_j [15].
- K-means. The model considers thread's text as "Document-Term"matrix. Each word is then assigned to its closest cluster center and the center of the cluster is updated until the state of no change in each cluster center is reached [31].
- Expert. Selecting keywords for each thread by experts. Choosing the most semantically meaningful words based on the thread topic.
- Latent Dirichlet Allocation (LDA). See the description above.

[1] http://www.ruscorpora.ru/en/index.html
[2] https://code.google.com/archive/p/word2vec/

3.4 Data collection

To collect our data, we used the following steps:

1. Select a forum and distinguish threads, which contain at least 400 posts.
2. Define user query. Mostly we try the query to be the same as thread name.
3. Collect all the posts from these threads with the following information: thread URL, post text, author, information about external sources in post.
4. Mark down sentiment value, informativeness and relevance of each post by criteria, listed in Table 1 .

The forums used in our work are listed in Table 4.

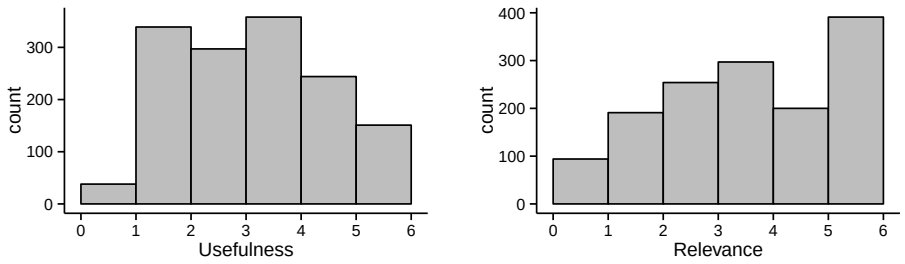
Table 4: The chosen Internet forums

Forum/URL	Thread title/Query
1 iXBT (Hardware forum) http://forum.ixbt.com/	Choosing of ADSL modem/How to choose ADSL router?
2 Fashion, style, health http://mail.figgery.com/	Diets for overweight people/How to lose weight?
3 Kinopoisk (cinema forum) http://forum.kinopoisk.ru/	"Sex at the city" series/How good is "Sex at the city" and why?
4 Housebuilding forum http://www.forumhouse.ru/	Building a house using 6x6 wooden planks/How to build a house using 6x6 wooden planks?
5 Velomania (bicycle forum) http://forum.velomania.ru/	Why are the pistons return to caliper?/Why are the pistons return to caliper?
6 Guitar players forum http://forum.velomania.ru/	All questions about guitar tuning/How to tune the guitar?
7 Evening dresses http://club.osinka.ru/	Wedding dresses/How to make the corset pattern?
8 Sewing the wedding http://thesewingforum. co.uk/	Wedding dresses/Dress for friends wedding - tips for sewing satin/How to handle a silk dress?

3.5 Experiments

For each selected machine learning model we do the following steps:

1. Split data into the train (70% of each forum) and test (30%) sets.
2. For the selected model, train set for each target variable and apply it to the test set of each collected forum.
3. Sort posts by decreasing target variable approximation and take the N top posts.
4. Calculate $NDCG$ for each selection using ground truth values for informativeness and relevance.



(a) Distribution of informativeness (b) Distribution of relevance

Fig. 1: Distribution of target variables

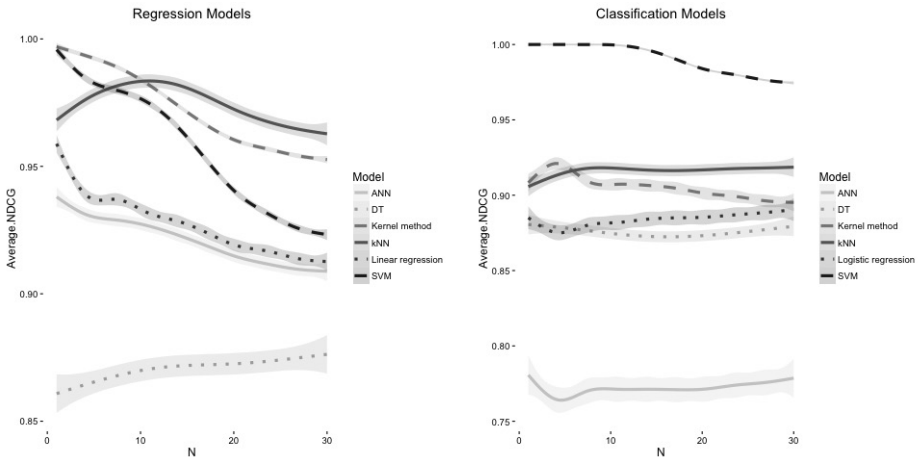
To ensure the model stability we used a bootstrap-like method. We resampled data for each iteration of steps 1–4, split it into train and test, fitted models and estimated quality. This process was repeated 200 times. Then we calculated the models average and its confidence intervals.

4 Results and Discussion

The Pearson correlation coefficient between informativeness and relevance on all forums is 0.36. This is an evidence of that these parameters are different, and query types expect IR system to do different things. Also, distribution of relevance is skewed towards 5 (see Fig. 1b), while the distribution of informativeness has the peak around 3 (see Fig. 1a). The skew of relevance is explained by the procedure of data collection: we choose posts from already relevant threads, so it is expected that most of the marked posts have high relevance. Distribution of informativeness shows that great portion of posts has moderate (2-3) informativeness, and only a small portion of posts have marginally high or low informativeness.

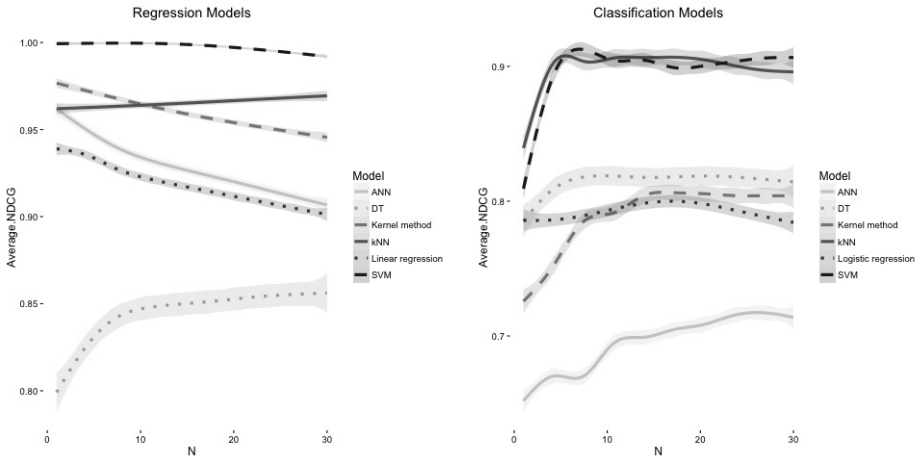
Fig. 2 and Fig. 3 shows the comparison of machine learning algorithms, that was listed in Table 3. There is not enough information there for selecting best algorithm for extracting professionally significant information. However, there is relatively high consistency (Kendall correlation coefficient was $\tau_6 = 0.73$) among the six studied algorithms. Fig. 4 shows the average quality for all algorithms of summarization for both target variables. As it can be seen, relevance is better described with regression methods, while in informativeness evaluation such dependence is weaker and backward. Also, as additional researches shows, the nature of nonlinearity depends on the specific features - there is a strong correlation between the length of the post and its informativeness (see Table 5).

Table 5 shows that keywords features are among first leading features in relevance evaluating. The Fig. 6 shows that informativeness is almost independent of keywords extraction method, while relevance is rather sensitive for its



(a) Informativeness, regression (b) Informativeness, classification

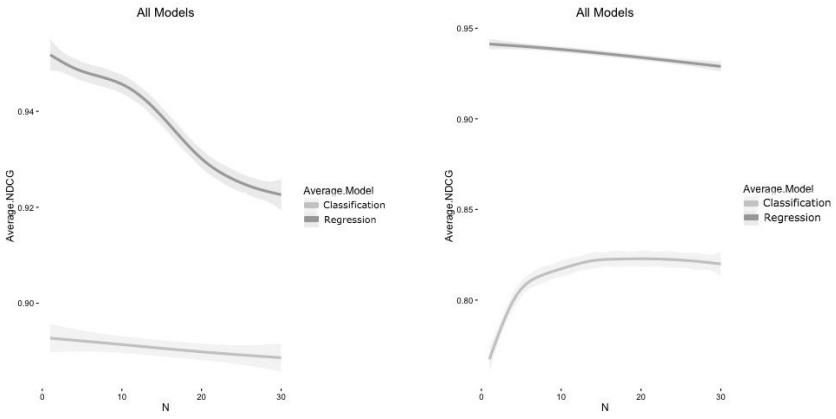
Fig. 2: Dependence of NDCG on informativeness and type of machine learning task type



(a) Relevance, regression (b) Relevance, classification

Fig. 3: Dependence of NDCG on relevance and machine learning task type

selection. Also, Most Used Keywords and Expert keywords are very similar in quality evaluation. Comparison of Fig. 5 and Fig. 6 shows that LDA approach and word2vec model show the worst performance. These models use only textual features. Therefore, non-textual features have great impact on summarization performance.

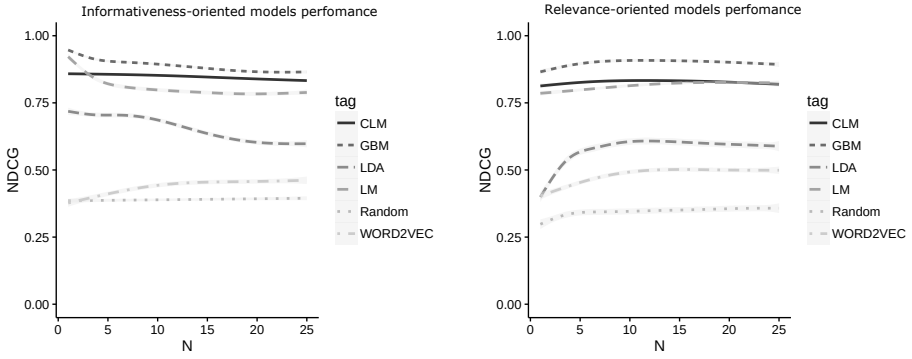


(a) Informativeness (b) Relevance

Fig. 4: Dependence of average NDCG on target variable and machine learning task type

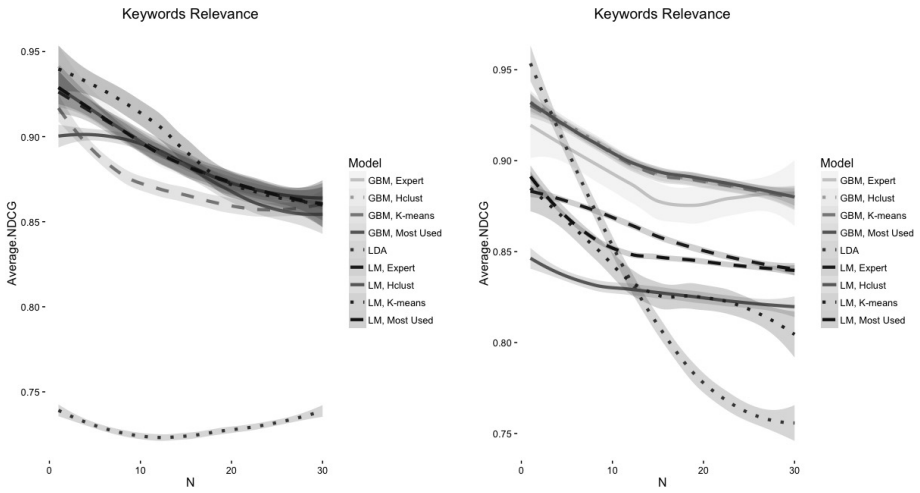
Table 5: Best Features

Algorithm	Multiple Linear Regression	Gradient Boosting Model
Target Variable	Relevance	Informativeness
Best Features	Query Keyword Count Most Used Keyword Count Author inDegree Author inDegree with Sentiment value	Length Author outDegree Author outDegree with Sentiment value Post position in thread



(a) Target variable informativeness (b) Target variable is relevance

Fig. 5: Dependence of NDCG on target variable and model type



(a) Distribution of informativeness (b) Distribution of relevance

Fig. 6: Comparison of keywords extraction methods

Regarding the results, our assumptions are the following: the relevance of posts is crucially determined by their lexical features, while the informativeness is related to the semantics of the forum in general and expressed in terms of the characteristics of the posts as a linguistic structure, as well as of the social graph of the forum. So the methods based on model "bag-of-words"such as classical search of keywords or topic modeling can be quite useful to highlight the relevant posts. At the same time, to extract informativeness posts, it makes sense to use specialized algorithms based on principals used in our work. The real interest to real information retrieval systems is the generalized extraction of information for the queries of different types.

5 Conclusion

In this work, we consider the problem of automatic summarization of professionally significant information from web forums. We have collected a big dataset, which contains threads from web forums about different topics. We showed, that the context of query plays an important role in evaluating of information extraction from forums. The informativeness-oriented and relevance-oriented queries are different by nature and have a weak correlation of their target variables. Relevance is best described by a linear combination of features. Also, the method of keywords extraction plays a big role in model effectiveness, when the target variable is relevance. However, at the same time informativeness is better described by the non-linear combination of features and depends on the social graph of the forum and overall textual structure of the thread. In our future work we want to investigate the practical use of the models, that were proposed in this paper.

Acknowledgement

Authors would like to thank Andrey Filchenkov for useful comments and proofreading. This work was financially supported by the Government of the Russian Federation, Grant 074-U01.

References

1. Agresti, A., Kateri, M.: Categorical data analysis. Springer (2011)
2. Al-Hashemi, R.: Text summarization extraction system (tses) using extracted keywords. Int. Arab J. e-Technol. 1(4), 164–168 (2010)
3. Almahy, I., Salim, N.: Web discussion summarization: Study review. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). pp. 649–656. Springer (2014)
4. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: An overview of graph-based keyword extraction methods and approaches. Journal of Information and Organizational Sciences 39(1), 1–20 (2015)
5. Bishop, C.M.: Pattern recognition. Machine Learning 128 (2006)
6. Biyani, P., Bhatia, S., Caragea, C., Mitra, P.: Using non-lexical features for identifying factual and opinionative threads in online forums. Knowledge-Based Systems 69, 170–178 (2014)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
8. Bottenberg, R.A., Ward, J.H.: Applied multiple linear regression. Tech. rep., DTIC Document (1963)
9. Elbedweihy, K.M., Wrigley, S.N., Clough, P., Ciravegna, F.: An overview of semantic search evaluation initiatives. Web Semantics: Science, Services and Agents on the World Wide Web 30, 82–105 (2015)
10. Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4), 367–378 (2002)
11. Grozin, V., Dobrenko, N., Gusarova, N., Ning, T.: The application of machine learning methods for analysis of text forums for creating learning objects. Computational linguistics and intellectual technologies 1, 199–209 (2015)
12. Grozin, V.A., Gusarova, N.F., Dobrenko, N.V.: Feature selection for language independent text forum summarization. In: International Conference on Knowledge Engineering and the Semantic Web. pp. 63–71. Springer (2015)
13. Harman, D.: Information retrieval evaluation. Synthesis Lectures on Information Concepts, Retrieval, and Services 3(2), 1–119 (2011)
14. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval 3(12), 1–224 (2009)
15. Lomakina, L., Rodionov, V., Surkova, A.: Hierarchical clustering of text documents. Automation and Remote Control 75(7), 1309–1315 (2014)
16. Lott, B.: Survey of keyword extraction techniques. UNM Education (2012)
17. Mikolov, T., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems (2013)
18. Nenkova, A., McKeown, K.: A survey of text summarization techniques. In: Mining text data, pp. 43–76. Springer (2012)
19. Nettleton, D.F.: Data mining of social networks represented as graphs. Computer Science Review 7, 1–34 (2013)

20. Oufaida, H., Nouali, O., Blache, P.: Minimum redundancy and maximum relevance for single and multi-document arabic text summarization. Journal of King Saud University-Computer and Information Sciences 26(4), 450–461 (2014)
21. Petrelli, D.: On the role of user-centred evaluation in the advancement of interactive information retrieval. Information processing & management 44(1), 22–38 (2008)
22. Ren, Z., Ma, J., Wang, S., Liu, Y.: Summarizing web forum threads based on a latent topic propagation process. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 879–884. ACM (2011)
23. Romero, C., López, M.I., Luna, J.M., Ventura, S.: Predicting students' final performance from participation in on-line discussion forums. Computers & Education 68, 458–472 (2013)
24. Saracevic, T.: Evaluation of evaluation in information retrieval. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 138–146. ACM (1995)
25. Schütze, H.: Introduction to information retrieval. In: Proceedings of the international communication of association for computing machinery conference (2008)
26. Sizov, G.: Extraction-based automatic summarization: Theoretical and empirical investigation of summarization techniques (2010)
27. Smine, B., Faiz, R., Desclés, J.P.: Relevant learning objects extraction based on semantic annotation. International Journal of Metadata, Semantics and Ontologies 8(1), 13–27 (2013)
28. Sondhi, P., Gupta, M., Zhai, C., Hockenmaier, J.: Shallow information extraction from medical forum data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 1158–1166. Association for Computational Linguistics (2010)
29. Tang, J., Yao, L., Chen, D.: Multi-topic based query-oriented summarization. In: SDM. vol. 9, pp. 1147–1158. SIAM (2009)
30. Wang, J.z., Yan, Z., Yang, L.T., Huang, B.x.: An approach to rank reviews by fusing and mining opinions based on review pertinence. Information Fusion 23, 3–15 (2015)
31. Wartena, C., Brussee, R.: Topic detection by clustering keywords. In: 2008 19th International Workshop on Database and Expert Systems Applications. pp. 54–58. IEEE (2008)
32. Zhao, H., Zeng, Q.: Micro-blog keyword extraction method based on graph model and semantic space. Journal of Multimedia 8(5), 611–617 (2013)