Comparison of Text Forum Summarization Depending on Query Type for Text Forums

Vladislav Grozin, Kseniya Buraya and Natalia Gusarova

Abstract Various approaches are developed for evaluation of query-oriented text summarization. However, for text forums this procedure is not well-defined, and standard approaches are not suitable. Evaluation of query-oriented text summarization greatly depends on the query type. We compare two typical scenarios of search of professionally significant information on Internet forums. Our subject of interest is the similarities and differences between relevance-oriented queries and usefulness-oriented queries. To compare these query types we have collected dataset, extracted textual, structural features and social graph features, constructed different ranking models, used suitable quality measure (NDCG), and applied feature selection techniques to investigate causes of differences. We have found out that these query types are very different by their nature, have weak correlation. Distinct model types and features should be used in order to create an efficient information retrieval system for each query type.

1 Introduction

Nowadays the value of professionally important information is rising steadily. Specialized web-forums are a valuable source of knowledge of that kind. Forums contain experience of people who actually used the technology and its features. Moreover, forums contain both positive and negative experiences—something that is not available from official documentation at all. But usually the majority of posts

V. Grozin (✉) · K. Buraya · N. Gusarova
Mechanics and Optics, National Research University of Information Technologies, Saint-Petersburg 197101, Russia
e-mail: 161397@niuitmo.ru

K. Buraya
e-mail: ks.buraya@gmail.com

N. Gusarova
e-mail: natfed@list.ru

P.J. Soh et al. (eds.), *Advances in Machine Learning and Signal Processing*, Lecture Notes in Electrical Engineering 387, DOI 10.1007/978-3-319-32213-1_24

at forums are useless and superfluous, containing a lot of hackneyed, repeated and irrelevant information. The obvious solution for this problem is using techniques of text summarization.

The text summarization is one of the tasks of information retrieval. It is about automatically extracting the main gist of the given documents to indicate the main aspects in them. This task is being actively investigated yielding a wide range of approaches, search mechanisms, results management and presentation (see, for example, [1]).

A crucial issue of the text summarization is evaluation problem, involving information retrieval effectiveness, or assessing consumers's satisfaction with the system [2]. Various approaches are developed for an assessment of text summarization. First of all, the approaches based on the 'bag of words' model are widely used. Typically the experimental queries are generated by extracting keywords from the list of terms frequently searched for within the field of interest (see, for example, [3]).

Besides, there is a set of search evaluation initiatives and competitions like TREC, DUC and MUC. They have created methodologies that can be conducted in a controlled lab-based setting. The most used is the Cranfield methodology [2] based on specialized test collection containing a set of predefined topics describing typical users' information needs.

Evaluation from a user-oriented perspective goes beyond the traditional Cranfield style experiments [2]. A common approach is to investigate users behavior in retrieval tasks in a controlled lab-based environment. Questions identified by the researcher are used here instead of predefined queries.

However, there is no track devoted to web-forums within the list of tracks managed by these evaluation initiatives [1].

Therefore, the paper deals with applying standardized evaluation approaches for text forums. Evaluation greatly depends on query type. TREC distinguishes two query types: usefulness and relevance-oriented. This paper is concerned with finding similarities and differences between these query types in order to find whether this approach is applicable for text forums.

2 Related Works

2.1 Terminology

There is a various terminology for information retrieval evaluation. Saracevic et al. [4] distinguish six levels of evaluation for information systems (including IR systems). The first three levels are referred to measuring system performance, the last three levels correspond to user-oriented evaluation. These may be assessed by different terms, including efficiency, utility, informativeness, usefulness, usability, satisfaction and the users search success [2].

The term relevance is vaguely used in literature. Some authors use it to refer to the degree of match between a document and a question; in contrast, other authors distinguished between relevance (similar to system relevance assessed by an external judge/expert) and pertinence (user-relevance assessed only by the real user with the information need represented in the question) (see the discussion in [2]).

Saracevic et al. [4] consider utility to be a more appropriate measure for evaluating information retrieval systems. A document has utility if it is pertinent (relevant as perceived by the user) and also contributes to the user knowledge in the context of the query (by providing information that was previously unknown). In our paper, we follow this opinion and adopt utility as a measure of usefulness and worth of the answers provided by the system to its users.

2.2 Methods of Forum Summarization

There are different approaches to the problem of text summarization. Main classifications are extraction-based and abstraction-based summarization as well as single-document and multi-document approaches. The majority of works in the area of forum summarization use extraction-based techniques and single-document approach [5]. Extractive forum summarization tasks are in turn divided into generic summarization (obtaining a generic summary or abstract of the whole thread) and query relevant summarization, sometimes called query-based summarization, which summarizes posts specific to a query [6].

The large variety of algorithms is used in both variants including naive Bayes classifier, statistical language models, topic modeling, graph-based algorithms etc. [3, 5–10]. In this paper we use algorithms of gradient boosting and linear regression which have already proved the efficiency for text forum summarization in our previous work [6, 7]. We also use for comparison a query-oriented algorithm based on LDA (see below for details).

2.3 Nearest Researches

We managed to find several researches with the aim close to our work in literature. Grozin et al. [7] consider reviews posted in web, assessing "Review Pertinence" as the correlation among review and its article. Tang et al. [8] consider the sentence relevance and redundancy within the summarized text. Their maximum coverage and minimum redundant (MCMR) text summarization system computes sentence relevance as its similarity to the document set vector. This idea is also used in [9] for cross-lingual multi-document summarization.

Some articles [10, 11] are devoted to comparing system effectiveness and user utility. Oufaida [10] compared traditional TREC procedure of batch evaluation and user searching on the same subject. Petrelli [11] confirmed that test collections and

their associated evaluation measures do predict user preferences across multiple information retrieval systems. They found that NDCG metric most effectively modeled user preferences.

To sum up there are no articles dedicated in deep details to the problem discussed in our article.

3 Experiment

Our goal is to create models that efficiently retrieve posts for different query types from text forums that will satisfy users needs, and investigate differences and similarities between query types. In our work, we examine two query types (and thus, construct two ranking model types):

- Query which target is to retrieve objective and interesting information in the domain of subject of interest (informativeness). This query type focuses on extracting pieces of information that contribute towards user's knowledge.
- Query which target is to retrieve any information related to the query (relevance). Text forum can contain posts that are relevant (related) to the query; the goal of this scenario is to fetch these posts.

Therefore, we have to study informativeness-oriented queries and relevance-oriented queries, their similarities and differences. Note that these post informativeness and relevance maybe be independent: posts can be irrelevant, yet informative (detailed explanation of something that is related to the domain of the query, but not related to the user query itself), and posts can be relevant, yet non-informative (thread-starting questions).

3.1 Data Collection

To collect our data, we used following algorithm:

1. Select a forum and a narrow user query within. The query is defined as a set of keywords.
2. Select some threads within the forum which titles contain query keywords. This is done to reduce amount of obviously non-informative and irrelevant posts, and reduce amount of required expert time.
3. Copy information about all the posts from these threads: post text, author, and thread URL.
4. Mark down sentiment value, informativeness and relevance of each post.

Formal criteria for marking up informativeness, Relevance and Sentiment are listed in Table 1. The forums used in our work are listed at Table 2. Each thread collected from forum contains at least 400 posts.

Table 1 Formal markup criteria

Parameter	Value	Comment
Informativeness	0	Post contains no useful information
	1	Post gives some useful information, but most of it is not useful
	2	Post gives some useful information, but it is
	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	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
Sentiment value	−2	Post contains clearly expressed negative emotions
	−1	Post contains humble negative emotions or sarcasm
	0	Post has neutral sentiment value
	1	Post is overall cheerful and contains signs of joy or happiness
	2	Post contains clearly expressed positive emotions and exaltation

Table 2 The chosen Internet forums

Forum		Query	URL
1	iXBT (hardware forum)	Choosing of ADSL modem	http://forum.ixbt.com/
2	Fashion, style, health	Diets for overweight people	http://mail.figgery.com/
3	Kinopoisk (cinema forum)	"Sex at the city" series	http://forum.kinopoisk.ru/
4	Housebuilding forum	Building a house using 6 × 6 wooden planks	https://www.forumhouse.ru/

3.2 Models and Parameters

We have to construct set of models to estimate informativeness and relevance. Two models were used to estimate each target parameter:

- Linear model. It is interpretable, and it captures linear dependencies well. We used non-regularized linear model.

- Gradient boosting model. It is interpretable, and it can capture nonlinear dependencies. We used three CV folds to estimate the best amount of trees; 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.
- LDA. This 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 presence of these keywords to estimate target variables. It is expected that these subsets will have different properties (for example, "offtopic" and "on-topic"). For hyperparameters we have chosen 100 iterations and 3 topics.

Models for each target variable were constructed independently, but using the same technique, same train and test sets, and same set of features for linear and gradient boosting models.

Despite the fact that our target variables have six discrete grades, we treated them as quasi-continuous and used models in regression mode to avoid sparse class population because we have multiple strictly ordered classes.

To fit models we divided the data from each forum into train (70 % of each forum) and test (30 %) sets. 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 is repeated 200 times, and model qualities are averaged and confidence interval is calculated.

3.3 Quality Estimation

Widely used recall/precision metrics are not useful in our context, because we have ordered multiple classes for each target variable. It is recommended to use cumulative gain metrics to evaluate retrieval system quality [2]. We used normalized cumulative gain. It is a cross-query comparable metric that lies between 0 and 1. It is calculated using formula:

$$NDCG_N = \frac{DCG_N}{IDCG_N} \tag{1}$$

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

where N is the size of resulting set (how many documents to retrieve), rel_i is true value of target variable (relevance or usefulness) of ith post in the retrieved set, and $IDCG_N$ is maximum possible DCG_N for specified forum and N, i.e. DCG_N for ideal algorithm. The full procedure of model quality estimation for both query types is:

1. Fit models to train set of each forum for each target variable (usefulness,relevance) and apply them to test set of each forum. This gives $Usefulness_{est}$ and $Relevance_{est}$, some approximation of true usefulness and relevance values of test set.
2. Sort posts by decreasing target variable approximation ($Usefulness_{est}$ or $Relevance_{est}$) and take N top posts. This gives selection of N best posts according to the model.
3. Calculate NDCG metric for the selection using true usefulness and relevance values of this N best posts subset.

We variated N from 2 to 30 to investigate how models behave in case of different selection windows.

3.4 Features

We have to extract features for linear and gradient boosting models that will hint us on how useful or relevant is the specific post. There are a lot of possible features we can extract; we used the ones that are suitable for our case. Chosen features are listed at Table 3.

Table 3 Features

Type	Feature	What this feature means
Post's author graph features	Betweenness, non-sentiment graph inDegree, non-sentiment graph outDegree, non-sentiment graph	Author's social importance
		How many times author was quoted
		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's 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's 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)	Post's usefulness
	Number of query keywords	Topic conformity
	Most used topic keyword count	Topic conformity

Sentiment value was marked down by experts and is used as a feature. It is expected that posts with a positive attitude will be more useful.

Also, simple non-semantic text features were extracted: text length in characters, number of links and number of keywords within text. We used two algorithms of keyword extraction. First one splits the query into words, and treats them as keywords. A more extensive list of keywords would mean a search for synonyms and equivalents; it requires semantic analysis and is not available for every language and for every query domain. The second algorithm creates frequency table for each thread, and takes top 5 most popular words. In both algorithms, stopwords were stripped.

We represented social structure in the form of a social graph, where the nodes are the users, and edges indicate a link between two users. For the creation of the social graph we have used citation analysis: if person A quotes person B by explicitly mentioning his name in text, there is a guaranteed connection between A and B. We used two methods: a non-sentiment graph (edge weight is always 1) and a sentiment graph (edge weight is related to the post's sentiment value). After the creation of the graph parallel edges' weights were summed. Then, the weights of the edges were inverted.

Node centrality is often used to find people who are important members of society. We considered some proven [12] metric to evaluate node centrality: Betweenness centrality—the number of shortest paths between all pairs of nodes that pass through the node; inDegree—the total weight of incoming edges; outDegree—the total weight of the outgoing edges.

Position in thread is calculated as position of post in chronological order (first post has position in thread equal to one, next post has value of two etc.).

4 Results and Discussion

Correlation between usefulness 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 distribution of usefulness has 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 marked posts have high relevance. Distribution of usefulness shows that great portion of posts has moderate (2–3) usefulness, and only a small portion of posts have marginally high or low usefulness.

Figure 2 shows result of application of the procedure described at Quality estimation section. Plotted lines have 99.5 % error bands.

As one can notice, linear model is better at selecting relevant posts, and gradient boosting model is better at estimating usefulness. This means that relevance can be better described as a linear combination of the features, and usefulness is best approximated as a non-linear construction over calculated features.

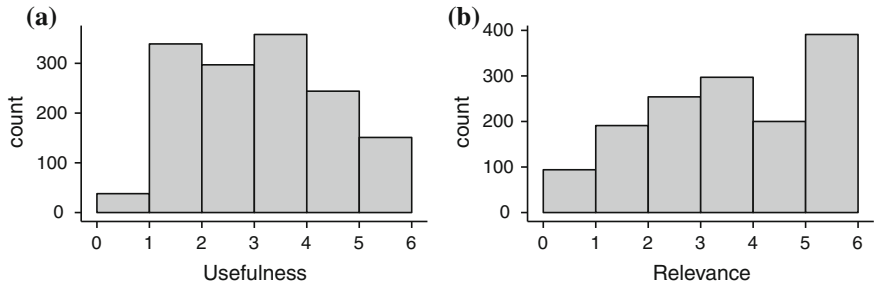

Fig. 1 Distribution of target variables. **a** Distribution of usefulness. **b** Distribution of relevance

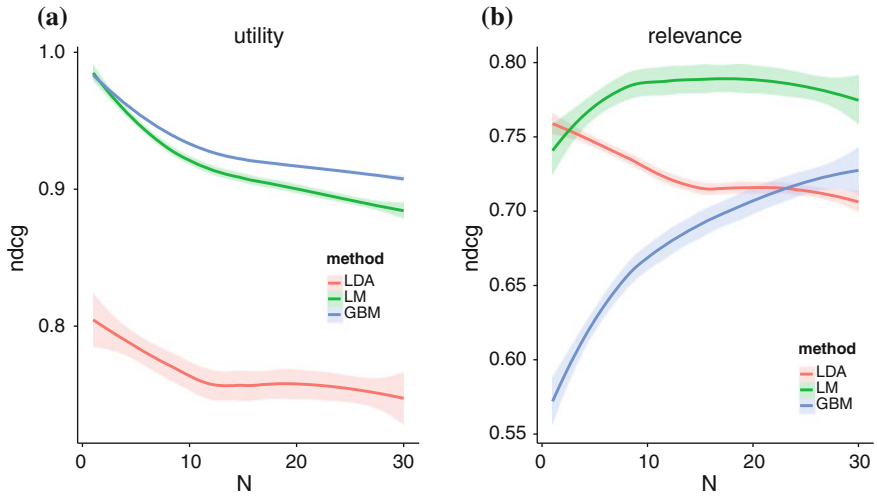

Fig. 2 Dependence of NDCG on target variable and model type. **a** Target variable is usefulness. **b** Target variable is relevance

For better comparison of query types, we have to investigate which features were best in each model. To do this we have chosen most important features (significance level of 0.001) from linear model constructed for relevance, and best features from gradient boosting model (the best models for each target variable). Feature selection from GBM was done by selecting top 4 features using relative influence metric [13]. The results are presented at Table 4.

Relevance is best estimated using keyword-related features, and usefulness is best estimated using post length and position in thread. Also, graph features appear in best feature list. This means that relevance-oriented are quite different from usefulness-oriented queries. Relevance-oriented queries can be handled by keyword-based features, and usefulness-oriented queries require simple textual and

Table 4 Best features

Model type	Linear regression	Gradient boosting model
Target variable	Relevance	Usefulness
Best features	– Query keyword count	– Length
	– Most used topic keyword count	– Author outDegree (social graph feature)
	– Author inDegree (social graph feature)	– Author outDegreeSent (social graph constructed using sentiment values feature)
	– Author inDegreeSent (social graph constructed using sentiment values feature)	– Post position in thread

structural features. Both model types can be improved by incorporating social graph features.

Note that despite the fact that relevance-oriented and usefulness-oriented queries are different types of queries that require different ranking methods, in real systems these models can be merged [14] in order to retrieve both relevant and useful posts.

5 Conclusion

We have defined query types to consider, collected dataset from four forums, constructed features and models, estimated model quality and interpreted the results to compare query types. The usefulness-oriented and relevance-oriented queries are different by nature, and have weak correlation of their target variables. Relevance-oriented queries are best handled using keywords-based features and linear model while usefulness-oriented queries are best handled using gradient boosting model and textual and structural features.

References

1. Nenkova A, McKeown K (2012) A survey of text summarization techniques. Min Text Data Springer 43–76
2. Elbedweihy Khadija M, Wrigley Stuart N, Clough Paul, Ciravegna Fabio (2015) An overview of semantic search evaluation initiatives. Web Semant Sci Serv Agents World Wide Web 30:82–105
3. Bhatia S, Mitra P (2010) Adopting inference networks for online thread retrieval. In: Proceedings of the twenty-fourth AAAI conference on artificial intelligence. Atlanta, Georgia, pp 1300–1305
4. Saracevic T, Kantor P, Chamis AY, Trivison D (1988) A study of information seeking and retrieving. I Backgr Methodol J Am Soc Inf Sci 39:161–176

5. Ren Zh, Ma J, Wang Sh, Liu Y (2011) Summarizing web forum threads based on a latent topic propagation process. In: CIKM11, October 2428. Glasgow, Scotland
6. Grozin VA, Gusarova NF, Dobrenko NV (2015) Feature selection for language-independent text forum summarization. In: International conference on knowledge engineering and semantic web, pp 63–71
7. Grozin VA, Dobrenko NV, Gusarova NF, Ning T (2015) The application of machine learning methods for analysis of text forums for creating learning objects. In: Computational linguistics and intellectual technologies, vol 1, pp 199–209
8. Tang J, Yao L, Chen D (2009) Multi-topic based query-oriented summarization. In: Society for industrial and applied mathematics—9th SIAM international conference on data mining 2009, Proceedings in applied mathematics 3, pp 1141–1152
9. Wang Jun-ze, Yan Zheng, Yang Laurence T (2015) Ben-xiong Huang An approach to rank reviews by fusing and mining opinions based on review pertinence. Inf Fusion 23:3–15
10. Oufaida Houda, Nouali Omar, Blache Philippe (2014) Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. J King Saud Univ Comput Inf Sci 26:450–461
11. Petrelli Daniela (2008) On the role of user-centred evaluation in the advancement of interactive information retrieval. Inf Process Manage 44:22–38
12. Borgatti Steve (2005) Centrality and network flow. Soc Netw 27(1):55–71
13. Friedman J (2001) Greedy boosting approximation: a gradient boosting machine. Ann Stat 29:1189–1232
14. Croft WB (2015) Combining approaches to information retrieval. In: Croft WB (ed) Advances in information retrieval. Kluwer Academic Publishers, Boston