Adaptive Discussion Forum for Reduce Information Overload

Aleksandr Kozko
Institute of Information Technologies
Chelyabinsk State University
Chelyabinsk, Russia
alkozko@yandex.ru

Andrei Melnikov
Institute of Information Technologies
Chelyabinsk State University
Chelyabinsk, Russia
mav@csu.ru

***Abstract* — nowadays an extremely large number of people is involved in the process of creating new information. This causes information overload, which reduces the effectiveness of traditional means of online interaction. We propose an approach that we hope will help to work with the increasing volumes of information. In this paper, we describe the algorithms and methods of forum restructuring aimed at simplifying the user's work and reducing the information overload for the user.**

Keywords — discussion forum, semantic analisys, information overload

I. Introduction

Today, more and more people interact with each other via various online means of communication due to the development of the Internet.

Discussion forums, blogs, comments, and question-and-answer systems are the most popular online means of mass communication nowadays, but these means were designed a long time ago. At that time, the forums had just hundreds of users, and everyone could easily navigate the forum discussions because the forum did not contain a lot of information.

Since then, the number of people who use the Internet and forums has significantly increased, they create a huge amount of information, and the volume of information is increasing every minute. Unfortunately, forum engines and other tools have not changed significantly since the time they were created, and today they no longer meet current requirements.

Currently millions of people all over the world use forums for various discussions - from entertaining conversations to a discussion of education and technology. At the same time, new users find it difficult to navigate the informational flow. This leads to a large number of duplicate topics, senseless messages and unanswered questions. All the above causes informational overload [1].

Forum engines have greatly evolved externally since the creation of the first forum, but their essence has remained almost the same. This indicates the need for more radical changes, which requires the use of intellectual means of information processing designed specifically for the forums.

II. Discussion Forum Structure

We single out four main means of online interaction: blogs, forums, question-and-answer systems and comment systems. It should be noted that all these means have a similar structure and differ only in ideology. Therefore, we can examine these means by the example of forums as the most common tools.

The forum structure is a tree, at the top of which there is a root page which contains the section. The sections, in turn, contain topics. Each topic consists of a root post and comments. Comments are presents in the form of a tree (explicitly or implicitly) and consequently form branches of the discussion.

Moderators and administrators create sections, users create posts and topics, and all this taken together forms a described structure.

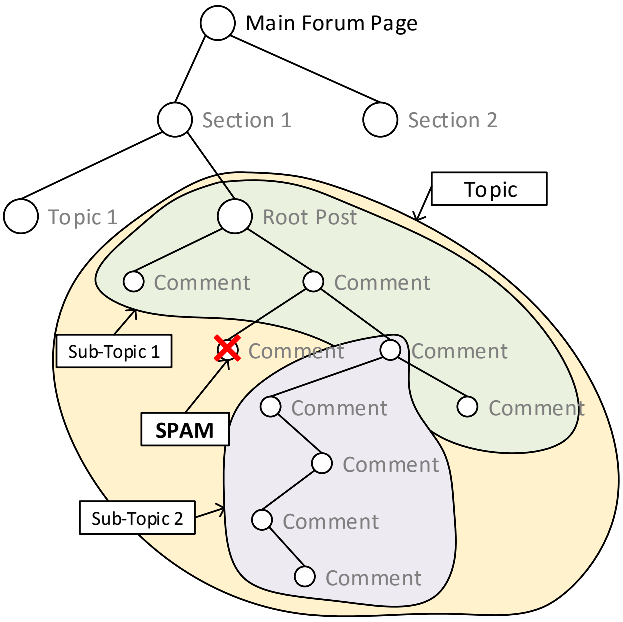

Fig. 1. Discussion Forum Structure

However, this structure is rather fixed and difficult to change, that is why it often does not correspond to the semantic content of the forum and its topics because each topic may contain meaningless, junk messages and subtopics – branches of discussions on the topic different from the one stated in the root post. This is clearly shown in figure 1.

All this leads to the informational overload, which implies the duplication of the forum content and separation of the discussion into several topics. This results in the difficulty in using forums as means of communication and collaboration for a large number of people.

III. Existing Methods for Reduce Information Overload

The attempts to solve the problem of information overload in online forums have been made long ago, but usually it is solved with the help of administrators and moderators. Paper [3] is devoted to the research of the crowdsourcing approach. Also, there have been attempts to solve this problem at the level of specific tools, e.g. with the help of tags. The study of the applicability of such tools is described in paper [4]. Some researchers attempt to apply qualitatively new methods, such as, for example, collaborative filtering [5], used for reduction of the overall information noise. However, there is no automated method of solving this problem that could be widely implemented.

IV. Adaptive Forums

In order to improve the effectiveness of the forums as a means of information exchange it is necessary to reconsider the concept of online forums in terms of adding the intellectual support means of online communication. We propose the concept of adaptive educational forums, described in the paper [6]. The main idea of this concept is the forums' ability to adapt itself for a user through changing forum structure in accordance with the semantic features of discussions and the information needs of the users.

Example of forum restructuring is clearly shown in figure 2.

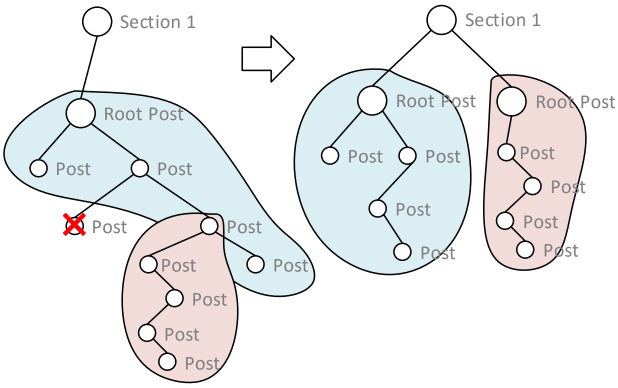

Fig. 2. Discussion of Forum Restructuring

However, to achieve this goal it is necessary to solve several tasks, one of which is to estimate the semantic connection between the posts in the forum to their further "bonding" into subtopics.

The next task is to develop the methods of forum restructuring

And the last task is software implementation of the system based on described methods and algorithms.

V. Semantic Proximity of Forum Posts

The measure of semantic proximity of the posts indicates how strongly the two posts are connected between themselves in meaning. We use this mesure to remove SPAM and junk posts and to distinguish subtopics – subset of discussion threads, which contain posts with semantic proximity united by one topic.

We propose the following algorithm to determine this metric:

1) To distinguish key terms

2) To determine the proximity between the key terms

3) To estimate the proximity of messages based on the similarity between the key terms in the message

A. Key Terms Distinguishing

In this work, the key term means a word or a phrase that describes the thematic focus of the document (a post in our case).

Almost all currently existing methods for distinguishing keywords work as follows [7]:

1. The normalization of an input document – refinement of extra symbols from the document, words normalization, etc.
2. Distinguishing of key terms candidates.
3. Filtering of the candidates by statistical and lexical features.
4. Assessment of the candidates by assigning weight to each of them.
5. Choice of key terms. This step may occur by threshold weight value or by a predetermined number of selected keywords.

Step 4 is the most interest step, because the choice of weight evaluation method affects the overall algorithm effectiveness. The most popular method for candidate's weight assignment is TF-IDF metrics and its modifications.

One of the main disadvantages of TF-IDF approach is that a set of documents should not be changed during the calculation. Adding a new document will require a recalculation of values. Several solutions have been proposed to solve this drawback, for example, the TF-ICF algorithm [8]. Mihalcea and Csomai evolved this approach in 2007 and described in their work a method using Wikipedia as a training set [9]. They used manually marked information from Wikipedia articles for assessing the weight of the candidate terms.

This assessment is rather accurate because the training is based on manually labeled Wikipedia database. However, it can be unreliable for seldom-used terms. Thus, the authors recommend to consider only those terms that appear in the Wikipedia at least 5 times.

In our view, the use of this approach is the most appropriate for the task of determination of keywords from the forums, because it is based on the manually labeled and updated article collection, which should provide good accuracy and relevance of assessment.

B. Assessment of Key Terms Proximity

There are many ways to define semantic proximity of two words that can be divided into [10]:

- methods based on ontologies: Reznik's measure, Leacock-Chodorow's measure and others;
- methods based on a corpus of texts: for example, Latent Semantic Analysis;
- methods based on Wikipedia and using the properties of partially structuring of the information in Wikipedia.

Forums have some features that must be considered when selecting a class of methods for determination of semantic proximity. These features include thematic focus of discussion, presence of specific terms in posts, and various subject areas of discussions.

Based on this, the usage of metrics based on the text corpus and ontologies is very complicated, due to the complexity of creation and maintaining of sufficiently complete ontology or text corpus. One of the ways to avoid this is use of Wikipedia as a corpus, because it constantly updates and covers the concept of the different subject areas. Wikipedia also contains meta-information and semi-structured data, which increases effectiveness of algorithms, due to it we propose to use one of wiki-based methods for determination of semantic proximity between terms.

Most wiki-based methods consider articles as concepts, and allow to determine the proximity between the concepts by using articles location in category tree or hyperlinks between the articles. However, this approach leads to appearance of the task of solving word-sense disambiguation for matching the input term and the wiki concept. Explicit Semantic Analysis (ESA) is a method that does not have this disadvantage and shows a good result on test sets [11]. We propose to use this method for determination of semantic proximity for discussion forum analysis.

C. Assessment of forum messages semantic proximity

Using data about keywords in the posts and the degree of proximity between them we can assess the semantic proximity between the two posts. We propose to calculate this value based on proximity of key-term pairs. We will consider two key-terms from two different posts as a pair of related key-terms, which have quite high value of proximity between them. If the value of proximity is less than the minimum threshold value, then such terms are not viewed as a pair and their bond is not considered in the overall assessment of posts connection.

As we propose to consider only the key terms in the message, each message can be represented as a set M = {k_1, k_2, …, k_n}, where k is a key term, we expect that forum message will contain usually one or two and no more than two or three key terms.

We can assess the semantic proximity between two forum messages basing on the semantic distance between key terms in the messages. For that, we offer to use inverted value (equation 1)

$$dist(k_1, k_2) = \frac{1}{sim(k_1, k_2)} \qquad (1)$$

where k_1 and k_2 are key terms.

Then we can calculate the semantic distance between two messages M_1, M_2, which are represented as *{k_{11}, k_{12},…, k_{1n}}* and *{k_{21}, k_{22}, …, k_{2n}},* as a composition of normalized sematic distances between key term pairs (equation 2).

$$dist(M_1, M_2) = \prod_i dist_{norm}(k_{1i}, k_{2i}) \qquad (2)$$

where *i* is a number of key term pairs.

In this case, we offer to consider only a pair of key terms, the distance between which is less than the threshold value. Its value is proposed to be determined experimentally.

In this case, the more similar pairs of terms is contained in the messages, the smaller distance between these messages is. Here each pair in messages improves the estimation of the distance between messages.

VI. Methods for Forum Adaptation

We propose the following group of methods based on the semantic content of the forum for automatic restructuring of the forum.

1) Concealment of senseless, junk and spam messages from threaded discussion through semantic analysis of each forum post.

2) Aggregation of the messages from one topic to several subtopics by semantic proximity metric.

3) Decomposition of forum topic into several topics through highlighting the subtopics and allocation of each subtopic into new topics.

4) Combination of related discussions from various topics into a new topic.

5) Transfer of topic to another section by classifying methods.

6) Creation of a new forum section by clustering methods.

All of these methods aim to rebuild the forum structure according to semantic content. Such adaptation is designed to simplify navigation and search of information and reduce of information overload, as we hope this approach will provide a better structuring of information.

VII. ConClusion

Today, massive online interaction tools face a huge amount of information that is created by users every day. Traditional tools, such as forums, have no instruments for effective work with huge information volumes, as a result it leads to information overload and reduce of forum effectiveness as a tool for cooperation.

In this paper we propose methods and algorithms for analysis of forum posts and change of forum structure according to forum content. We hope that the proposed approach will help to improve the efficiency of work with information on the forum and will reduce the information overload effect.

References

[1] J. Kim, “Influence of group size on students' participation in online discussion forums” in Computers & Education, 2013 Mar 31;62:123-9.

[2] C. Lampe, P. Zube, J. Lee, C.H. Park, E. Johnston, “Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums”. in Government Information Quarterly. 2014 Apr 30;31(2):317-26.

[3] G. Simarpreet, “Exploring the benefits of tagging forum posts based on a hierarchical domain model of the course content in online forums.” PhD diss., 2014.

[4] P.Y. Wang, H.C. Yang, “Using collaborative filtering to support college students’ use of online forum for English learning”, in Computers & Education. 2012 Sep 30;59(2):628-37.

[5] A. A. Kozko “Construction of Adaptive Educational Forums Based on Intellectual Analysis of Structural and Semantics Features of Messages”, in *Supplementary Proceedings of the 4th International Conference on Analysis of Images, Social Networks and Texts (AIST'2015),* Yekaterinburg, Russia, April 9-11, 2015, pp. 46-51.

[6] S. O. Sheremetieva, P. G. Osminin, “On Methods and Models of Keyword Automatic Extraction”, *Bulletin of the South Ural state university. Series: Linguistics,* vol. 1(12), 2015

[7] J. W. Reed, Y. Jiao, T. E. Potok, B. A. Klump, M. T. Elmore, A. R. Hurson “TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams”, in *Proc. Machine Learning and Applications*, *ICMLA '06*, pp. 258-263, 2006.

[8] R. Mihalcea, A. Csomai, “Wikify!: linking documents to encyclopedic knowledge” in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, New York, USA, 2007, pp. 233-242.

[9] A. Panchenko “Similarity measures for semantic relation extraction”, PhD thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium, 2013.

[10] E. Gabrilovich and S. Markovitch “Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis.” in *IJCAI*, vol. 7. 2007, pp. 1606-1611.