ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ТЕКСТОВЫХ ФОРУМОВ ДЛЯ ПОДГОТОВКИ УЧЕБНЫХ ОБЪЕКТОВ

Грозин В. А. (vlad.grozin@yandex.ru),
Добренко Н. В. (graziokisa@gmail.com),
Гусарова Н. Ф. (natfed@list.ru)

Университет ИТМО, Санкт-Петербург, Россия

Нин Тао (603096136@qq.com)

Чанчуньский научно-технологический университет, Чанчунь, Китай

Ключевые слова: отбор признаков, глубинный анализ текстов, учебные объекты, текстовый форум, машинное обучение

THE APPLICATION OF MACHINE LEARNING METHODS FOR ANALYSIS OF TEXT FORUMS FOR CREATING LEARNING OBJECTS

Grozin V. A. (vlad.grozin@yandex.ru),
Dobrenko N. V. (graziokisa@gmail.com),
Gusarova N. F. (natfed@list.ru)

ITMO University, Saint Petersburg, Russia

Ning Tao (603096136@qq.com)

Changchun University of Science and Technology, Changchun, China

Nowadays the concept of a learning object (LO) is widely used in preparation of educational materials. Usually, LOs are parts or fragments of previously created educational content, which is very informative and pedagogically focused. However, concerning high-dynamic branches of science and technologies LOs tend to become outdated and trivial thus losing their educative value. In this situation, specialized text forums become a valuable source of knowledge. Forums contain experience of people who actually used the technology and its features. They contain both positive and negative experience—something that is not available from official documentation at all. However, they also contain many trivial, repeated and still irrelevant posts. Also, an expert needs to extract useful messages from text forums according to his individual learning objectives.

The paper deals with the task of automatically identifying texts potentially useful for preparation of textual educational materials within text forums. For our experiments, we have selected highly inflective languages with complex grammar and rather weak text analysis tools: French, German, Russian and Chinese (Mandarin). We have overviewed non-semantic text and social features of a text forum which indicate the suitability for creation of a textual LO. We have analyzed those features. For this purpose, we have constructed linear and non-linear models of machine learning and conducted feature selection. Even for the forums providing little information about chosen topics and forums with a lot of off-topic text in dataset, these models were better than the baseline selection methods.

Keywords: feature selection, text mining, learning object, text forum, machine learning

1. Introduction

Nowadays we are facing the rapid growth of the amount of information available online, so it becomes more difficult to organize educational process according to this growth. Besides, much more persons are being involved into educational process: they are not only students and teachers, but also instructors, self-taught learners and so on, each having his/her own educational goals. They search for educational materials that satisfy their own needs.

The concept of learning object (LO) is developed for this purpose. There are different definitions for LO, the most common are [IEEE (2002), Wiley D. A., ed. (2001)]. Namely, IEEE defines LO as 'any entity, digital or non-digital, that may be used for learning, education or training'. Wiley defines LO as 'any digital resource that can be reused to support learning'.

Usually, LOs are chunks or slices of previously created educational content. Authors [Griffiths J., Stubbs G., Watkins M. (2007)] offer to develop LOs using textual course materials as a basis or source and dividing it downwards until the smallest item of information or idea is reached. Raw assets that have no inherent pedagogical aim can also be considered as lower-level LOs [Boyle, T. (2003)]

But LOs concerning high-dynamic branches of science and technologies tend to become outdated and trivial thus loosing educativeness (defined as a property that reflects the educative value of a document) [Hassan S., Mihalcea R. (2009)]. So, when seeking educational information about new technology it is often useless to refer to existing collections of LOs. On the other hand, searching the Web using one of the current search engines frequently lead to the results which badly meet the requirements of educativeness.

In this situation specialized text forums become a valuable source of knowledge. Forums contain experience of people who actually used the technology and its features. They contain both positive and negative experience—something that is not available from official documentation at all. But they also contain a lot of trivial, repeated and still irrelevant posts. Also, expert needs to extract useful messages from text forums according to his/her individual learning objectives.

The obvious solution is to use techniques of text mining, for example text summarization or question answering systems. But the task of preparing LOs for high-dynamic branches of science and technologies has the specifics. Typical expert's question is "Is there any new (or unknown or interesting for my audience) information in this piece of text?" It is obvious that such form of a question is difficult for question answering systems. Usually, novelty consists in emergence of concepts and relations not known before; respectively the search query has to be rather wide (coarse-grained) in order to include them. It complicates application of semantic methods in text summarization. And, last but not least, the procedure of preparing LOs has to be simple and language-independent in order to be used by ordinary teachers.

In this paper, we address the task of automatically identifying information potentially useful for preparing educational materials (learning objects) within technical text forums. Specifically, we considered the non-semantic text and graph features that indicate suitability for creating textual LO (related to the chosen topic and containing detailed argumentation). Also, we examined dependence of these features' quality from forum language.

2. Related works

Information retrieval in our situation can be considered as a variant of educational data mining. It is known as a new growing research community since 1995, and different data mining techniques can be applied here [Romero C., Ventura S. (2007)]. This section reviews the related works from different dimensions: the works aiming to handle text information in online discussion boards (or forums) as well as the approaches of question answering systems.

The task of information retrieval from text forums is usually interpreted as Web Forum Thread Summarization and typically aims to give a brief statement of each thread that involving multiple dynamic topics. Traditional summarization methods are cramped here by some challenges [Ren et al. (2011)]. The first is topic drifting: as the post conversation progresses, the semantic divergence among subtopics will be widened. Besides, most posts are composed of short and elliptical messages, their language is highly informal and noisy, and traditional text representation methods have sufficient limitations here.

According to the survey in [Ren et al. (2011)], the majority of works in the area of forum summarization use extraction-based techniques [Spärck Jones K. (2007)] and single-document approach. A lot of research on automatic dialogue summarization use corpus-based and knowledge-based methods. For example, authors [Zhou L., Hovy E. (2005)] identify clusters in internet relay chats and then employ lexical and structural features to summarize each cluster. Authors [Ren et al. (2011)] propose a forum summarization algorithm that models the reply structures in a discussion thread. In order to represent information of online forum in a learning environment authors [Carbonaro A. (2010)] uses concept-based summarization: each word in the document is labeled as a part of speech in grammar, and to handle the word sense disambiguation problem similarity measures based on WordNet is used.

Statistical methods of dialogue summarization are also of great interest for preparing LOs. For example, in [Wang U., Cardie C. (2011)] unsupervised (TF-IDF and LDA topic modeling) and supervised clustering procedures (using SVMs and MaxEnt) are used in combination for decision summarization for spoken meetings. Authors [Biyani et al. (2014)] consider the problem of extracting relevant posts from a discussion thread as a binary classification problem.

There is a number of the works devoted to multi-lingual aspects of text summarization. For example, in order to fulfill sentiment analysis of multi-lingual Web resource authors [Hogenboom A. et al. (2014)] consider English as basic and use language-specific semantic lexicons of sentiment-carrying words. Contrary to this approach, authors [Banea C., Mihalcea R., Wiebe J. (2014)] show that the multilingual model consistently outperforms the cross-lingual one. Practical experience of developing natural language processing applications for many languages is described in [Steinberger R. (2011)]. The author considers Machine Learning methods as an extremely promising approach to develop highly multilingual systems.

A fast-growing number of studies have shown that the social factor can be useful in text forum summarization regarding to educativeness. For example, authors [Li Y., Liao T., Lai Ch. (2012)] apply similar measures as used in blogs to the forums, such as counting the number of common tags and replying or citing the same threads. Authors [Yang S. J. H., Chen I. Y. L. (2008)] explain that in an online forum context a central core (strongly connected component) contains users that frequently help each other by following questioner (requester)–answerer (expert) links.

A lot of Question Answering Systems are presented in literature (see the overview [Kolomiyets O., Moens M.-F. (2011)]). They differ by models and techniques depending on the system requirements, the type of question posed, the type of interrogated data, the type of interface and other criteria. In technical forums the conversation often inquires the solution to a specific problem faced by the user and others answer by adding their experience in that field [Almahy I., Salim N. (2013)]. If answering this question is the educational objective of the LO then the approach of Question Answering Systems can be useful.

But for complex questions a deeper semantic analysis is required (broader coverage of expected answer types, semantic role labeling and discourse analysis). For example, authors [Ferrandez et al. (2009)] solve Cross-Lingual Question Answering tasks. They make a syntactic analysis of the question using a shallow parser tool and extracting the syntactic blocks of a question. Authors [Cao et al. (2011)] process on-line forums at which questions are presented in an obvious form. They aim is extract contexts and answers for them and use the structural Support Vector Machine method. However, as the analysis of literature have shown, questions of the type, which is declared in Introduction aren't processed in known Question Answering Systems.

3. Methods

For our experiments we have selected some highly inflective languages with complex grammar and rather weak text analysis tools, in particular four languages—German, French, Russian and Chinese (Mandarin). Detailed information about the forums is presented in Table 1. From each forum we allocated threads which names contained a topic of interest used in the form of a keyword. These posts' usefulness for the creation of a textual LO with an appropriate educational value was manually marked down by experts (Table 2). We have invited experts of the relevant field who are native speakers in the languages of the forums.

Table 1

#	Forum	Language	Topic	Threads/ posts	Keywords
1	gamedev.ru	Russian	Unity	10/410	unity
2	hifi-forum.de	German	Windows vs Linux	13/173	windows, linux
3	forum.modelsworld.ru	Russian	Ship modeling	3/150	ship, model
4	5500.forumactif.org	French	Ship modeling	3/150	ship, model
5	bbs.csdn.net	Chinese	cocos2d-x	11/120	cocos
6	bbs.chinaunix.net	Chinese	Linux for beginners	11/103	linux

Table 2

Scale	Comment
0	Offtopic
1	Post is on the chosen topic, but argumentation is incomplete or absent
2	Post is on the chosen topic, and the author's point of view is well-argumented with explanations or external links

Nowadays there are a lot of works proposing different features for text forums, potentially suitable for educational value evaluation [Hassan S., Mihalcea R. (2009); Biyani et al. (2012); Smine et al. (2013); Dringus, Ellis (2005); Romero et al. (2013)]. However, not all of them are suitable for machine learning due to the specifics of our task. The list of the selected characteristics is presented in Table 3. In general, the calculation procedures were created using the sources mentioned above, but with the following specifics.

We calculated text sentiment value using sentiment keywords, specific for the forum's language. The resulting values were normalized to the range from -1 (strongly negative text) to +1 (strongly positive text).

Also, simple non-semantic text features were extracted: text length, number of links and number of keywords. Keywords were chosen strictly corresponding to the name of the forum topic. A more extensive list of keywords would mean a search for synonyms and equivalents, which requires semantic analysis.

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 [Tore Opsahl (2014)].

Node centrality is often used to find people who are important members of society. We considered some proven [Freeman L. C. (1978); White D. R., Borgatti S. P. (1994)] 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 number of post in chronological order (first post has position in thread equal to one, next one is equal to two etc.).

The features selected for machine learning are listed in Table 3.

Table 3

Type	Feature	What this feature means
Post's 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'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 feature	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 keywords	Topic conformity

The analysis was fulfilled using machine learning methods. Model creation was made in R. We chose two models: gradient boosting model in "gbm v.2.1" package [gbm package (2014)] (for base learners we chose trees and default model parameters were used; to determine the optimal forest size we used built-in cross-validation with three folds) and linear regression (lm). Each model was trained on training dataset (so, each forum had independent models). Each model was trained on a training dataset (so, each forum had independent models). Then, models predicted grades of each message in the test set, and N most qualitative candidates were selected. After this metric NCG metric was calculated and averaged among forums. Datasets for each forum were randomly divided into training (60%) and test (40%) sets.

To calculate selection quality we used Normalized Cumulative Gain metric [Järvelin K., Kekäläinen J. (2002)]:

$$NCG = \frac{\sum_{i=1}^{N} rel_i}{NCG_{\max}^{N}}$$

Where N is number of selected posts, rel(i) is quality of i-th selected post, and is maximum possible NCG for specified N. This metric lies between 0 and 1 (assuming rel(i) is non-negative) and indicates the quality of the selection, but this metric doesn't penalize late relevant items. It is assumed that the expert will still read all selected messages, so the order does not matter.

Following methods of post selection were chosen for baselines:

- Baseline-1: Head posts in thread have special importance and often contain more useful information (due to off-topic content in later messages) [Said D., Wanas N. (2011)]. So, one method is based on selecting the first messages of each thread.
- Baseline-2: Other method is using semantic keywords list [Steinberger R. (2011)]. A broad list of topic-related keywords and synonyms in English and in the forum language were made by experts. This method selects posts with the highest number of these words. Stemming and lemmatization (package "tm" in R) were used where possible.

4. Results and discussion

Fig. 1 shows the dependence of selection method (NCG metric) on the model and N.

As one can see, linear model (lm) was the best and both models were better than both baselines. So, because our smallest forum contains 100 messages (and test set has 40), we evaluated metrics for N varying from 1 to 30.

Since our both models were better than baselines, there are good features indicating suitability for making textual LO. Our ultimate goal is to investigate which ones. For this purpose we used feature selection methods.

For analyzing linear dependencies we used significance of features. The significance is probability of observing the data assuming linearly independence of regressor and explainable variable [Gelman A., Hill J. (2006)]. If this probability is less than a certain threshold (in our case—0.05), we can reject that hypothesis and say that there is dependence between variables.

Gradient boosting model was used to analyze non-linear dependencies. After training the model (gbm is a set of trees) we can see how many times trees were divided by each variables, and estimate split efficiency. This way we will get relative information influence metric, which can be interpreted as the importance of features [Gradient (2014).].

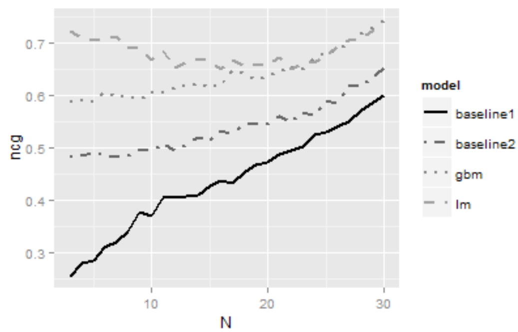

Fig. 1

Table 4 contains feature list along with their significance (S) and information importance metric (IMM) for each forum. Significance with less than 0.05 or non-zero IIM are highlighted with yellow background. Selected features (which had low significance or non-zero IIM at least once in every language) are also marked yellow. Of course, we need to consider statistical characteristics of the collected samples while selecting the features. For example: graph characteristics for forums 3 and 4 are invalid because of the small sample size and due to forum engine specifics (author names are not mentioned while quoting). Features which could not be calculated were marked as "N/A".

Table 4

Feature	#1		#2		#3		#4		#5		#6	
	S	IIM	S	IIM	S	IIM	S	IIM	S	IIM	S	IIM
Betweenness, non-sentiment graph	**0.04**	**1**	0.80	**1**	0.55	0	N/A	0	N/A	0	N/A	0
inDegree, non-sentiment graph	0.67	0	0.51	**13**	0.88	**0.7**	N/A	0	N/A	0	N/A	0
outDegree, non-sentiment graph	0.24	0	0.19	**1**	0.9	0	N/A	0	N/A	0	N/A	0
Betweenness, sentiment graph	0.17	0	0.80	0	0.57	0	N/A	0	N/A	0	N/A	0
inDegree, sentiment graph	0.67	0	0.51	**4**	0.1	**0.7**	N/A	0	N/A	0	N/A	0
outDegree, sentiment graph	0.26	0	**0.04**	**5**	0.91	0	N/A	0	N/A	0	N/A	0

Feature	#1		#2		#3		#4		#5		#6	
	S	IIM	S	IIM	S	IIM	S	IIM	S	IIM	S	IIM
Number of threads author is participating in	0.87	0	0.52	**4**	0.55	13.6	0.59	**17**	N/A	0	N/A	0
Position in thread	**0.02**	4	**0.04**	12	0.08	**54**	**0.04**	54	0.25	0	0.12	**14**
Times quoted	0.97	0	0.64	**6**	0.10	**1.6**	N/A	0	N/A	0	N/A	0
Length	**9e-8**	**80**	**0.03**	42	0.26	**15**	0.21	**49**	**0.001**	**44**	0.9	**24**
Links	0.53	0	0.97	0	0.47	**3**	0.71	0	N/A	0	0.49	0
Sentiment value	0.23	**2**	0.606	**5**	**0.001**	**22**	0.59	**39**	**1e-6**	**55**	**e-7**	**61**
Number of keywords	0.02	11	0.82	8	0.9	0	0.9	0	0.01	0	0.73	0

Posts with high educational value tend to have a positive sentiment. However, on forums ##3, 4 it has much higher significance than on forums ##1, 2, 5, 6. To explain this effect we have plotted the distribution of the number of messages from the Utility and Sentiment rounded to nearest ½. Fig 2 shows this distribution for forums ## 1, 2, 5, 6 (A) and forums ##3, 4 (B).

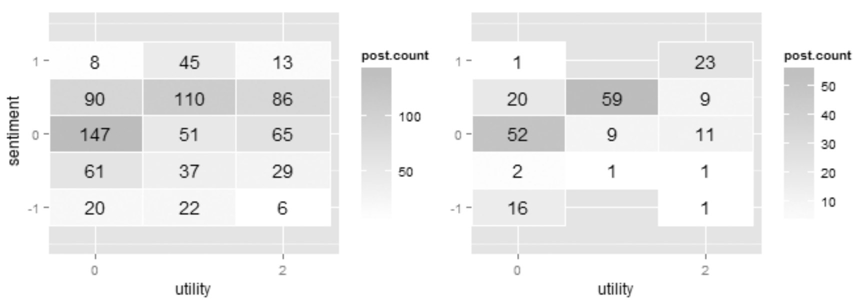

Fig. 2 (A, B)

As one can see, the distribution in the first picture (A) is nearly symmetrical with respect to sentiment=0, while the distribution in the second picture (B) is much more "skewed". For example, in fig. 2(A) value of post.count for sentiment=0.5 almost doesn't depend on value of utility. In other hand, in fig. 2(B) this dependence is obviously expressed, so a post from form B with sentiment equal to 0.5 is quite likely to have utility equal to 1. So, significance of sentiment is much higher on these forums. This "skewness" is associated with the strict moderation of forums (negative and offtopic posts are getting deleted, so users tend to leave useful and friendly texts). This is confirmed by the additional semantic analysis of the forum content, conducted by an expert.

Thus, we have selected features that indicate the potential educativeness of the post, and which are independent from the semantics of the post (see. Table 2). These are: the Length, Position in thread and Sentiment value.

5. Conclusion

In this paper, we have addressed the task of automatically identifying information potentially useful for the preparation of the educational materials (in particular learning objects learning objects) within technical text forums.

We have overviewed non-semantic text features that indicate suitability for creating LO (relating to chosen topic and containing detailed information), as well as social features from a text forum. The quality of the features was analyzed. Also, linear and non-linear models were constructed. These models were better than baseline selection methods even for forums with small samples on chosen topics and forums with a lot of off-topic text in dataset.

References

1. *Almahy I., Salim N.* (2013), Web Discussion Summarization: Study Review. Proc. of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Ed. Herawan T. et al. Springer Verlag, 2013. Pp. 649–658.
2. *Banea C., Mihalcea R., Wiebe J.* (2014), Sense-level subjectivity in a multilingual setting. Computer Speech and Language, Vol. 28, pp. 7–19.
3. *Biyani* et al. (2012) Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, Prasenjit Mitra. Thread specific features are helpful for identifying subjectivity orientation of online forum threads. Proc. of the COLING 2012, 24th International Conference on Computational Linguistics, Conference: Technical Papers, 8–15 December 2012, Mumbai, India, pp. 295–310.
4. *Biyani* et al. (2014), Biyani P., Bhati S., Caragea C., Mitra P. Using non-lexical features for identifying factual and opinionative threads in online forums. Knowledge-Based Systems, V. 69, October 2014, pp. 170–178
5. *Boyle, T.* (2003), Developing and delivering learning objects from a practitioner's point of view LTSN, available at: Generic Centre www.heacademy.ac.uk/resouces
6. *Cao* et al. (2011) Yunbo Cao, Wen-Yun Yang, Chin-Yew Lin, Yong Yu. A structural support vector method for extracting contexts and answers of questions from online forums. Information Processing and Management, Vol. 47, pp. 886–898
7. *Carbonaro A.* (2010), WordNet-based Summarization to Enhance Learning Interaction Tutoring. Peer Reviewed Papers. Vol. 6, n. 2, May 2010.
8. *Dringus, Ellis* (2005) Laurie P. Dringus, Timothy Ellis. Using data mining as a strategy for assessing asynchronous discussion forums // Computers & Education, Vol. 45, pp. 141–160
9. *Ferrandez* et al. (2009) Sergio Ferrandez, Antonio Toral, Oscar Ferrandez, Antonio Ferrandez, Rafael Munoz. Exploiting Wikipedia and EuroWordNet to solve Cross-Lingual Question Answering. Information Sciences, Vol. 179, pp. 3473–3488
10. *Freeman L. C.* (1978). Centrality in social networks: Conceptual clarification. Social Networks 1, pp. 215–239.
11. *gbm package* (2014), gbm: Generalized Boosted Regression Models. Available at: http://cran.r-project.org/web/packages/gbm/index.html

12. *Gelman A., Hill J.* (2006), Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, 2006. ISBN 978-0-521-68689-1
13. *Gradient* boosting machines, a tutorial. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/
14. *Griffiths J., Stubbs G., Watkins M.* (2007), From course notes to granules: A guide to deriving Learning Object components. Computers in Human, Vol. 23, pp. 2696–2720
15. *Hassan S., Mihalcea R.* (2009), Learning to Identify Educational Materials. International Conference RANLP 2009—Borovets, Bulgaria, pp. 123–127
16. *Hogenboom A.* et al. (2014), Hogenboom A., Heerschop B., Frasincar F., Kaymak U., de Jong F. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Systems, Vol. 62, pp. 43–53.
17. *IEEE* (2002), IEEE 1484.12.1–2002, 15 July 2002, Draft Standard for Learning Object Metadata, IEEE Learning Technology Standards Committee (LTSC)
18. *Järvelin K., Kekäläinen J.* (2002), Cumulated GainBased Evaluation of IR Techniques, ACM Transactions on Information Systems, Vol. 20, No. 4, pp. 422–446.
19. *Kolomiyets O., Moens M.-F.* (2011), A survey on question answering technology from an information retrieval perspective. Information Sciences, Vol. 181, pp. 5412–5434.
20. *Li Y., Liao T., Lai Ch.* (2012), A social recommender mechanism for improving knowledge sharing in online forums. Information Processing and Management, Vol. 48, pp. 978–994
21. *Owczarzak Dang* (2011) Karolina Owczarzak and Hoa Trang Dang. Who wrote What Where: Analyzing the content of human and automatic summaries. Proc. of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, June 23, 2011. Pp. 25–32.
22. *Ren* et al. (2011), Ren Zh., Ma J., Wang Sh. and Liu Y. Summarizing Web Forum Threads based on a Latent Topic Propagation Process. CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
23. *Romero C., Ventura S.* (2007), Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, Vol. 33, pp. 135–146
24. *Romero* et al. (2013) Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, Sebastián Ventura. Predicting students' final performance from participation in on-line discussion forums. Computers & Education, Vol. 68, pp. 458–472
25. *Said D., Wanas N.* (2011) Clustering posts in online discussion forum threads. International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No 2
26. *Smine* et al. (2013) B. Smine, R. Faiz and J-P. Desclés. Relevant learning objects extraction based on semantic annotation. Int. J. Metadata, Semantics and Ontologies, Vol. 8, No. 1, pp. 13–27
27. *Spärck Jones K.* (2007), Automatic summarising: the state of the art. Information Processing and Management, Special Issue on Automatic Summarising, 2007.
28. *Steinberger R.* (2011), A survey of methods to ease the development of highly multilingual text mining. applications. Language Recource and Evaluation, pp. 1–22
29. *Tore Opsahl* (2014). Node Centrality in Weighted Networks. Available at: http://toreopsahl.com/tnet/weighted-networks/node-centrality/

30. *Wang U., Cardie C.* (2011), Summarizing Decisions in Spoken Meetings. Proc. of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pp. 16–24, Portland, Oregon, June 23, c.
31. *White D. R., Borgatti S. P.* (1994) Betweenness centrality measures for directed graphs/ Social Networks, Vol. 16, pp. 335–346
32. *Wiley D. A., ed.* (2001), Connecting learning objects to instructional design theory: definition, a metaphor, and taxonomy. New York: Agency for Instructional Technology and the Association for Educational Communications and Technology, 2001
33. *Zhou L., Hovy E.* (2005), Digesting virtual "geek" culture: The summarization of technical internet relay chats. Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pp. 298–305, Stroudsburg, PA, USA.