

Всероссийский научно-практический симпозиум
**«СОЦИАЛЬНЫЕ КОММУНИКАЦИИ:
УНИВЕРСУМ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ»**

Санкт-Петербург, 9-10 ноября 2011 г.

Кольцова Олеся Юрьевна
Ефимова Татьяна Геннадьевна

**Выявление социальных проблем и изменений через анализ больших массивов
текстов в блогах и социальных сетях.**

Вот уже почти десятилетие как у социальной реальности появился новый «дом» - Интернет, ставший индикатором, предиктором и двигателем социальных процессов и изменений. «Арабские революции» этого года показали его значение даже в тех обществах, где доля пользователей Интернета не очень велика. Казалось бы, перед социологами открылись небывалые возможности для исследования и прогнозирования социальных процессов, однако исследовательское сообщество сталкивается с проблемой отсутствия средств анализа громадных объемов текстовых данных, который невозможен без специализированного программного обеспечения, продвинутых математических алгоритмов и компьютерно-лингвистических подходов. Такие социологические исследования оказались возможны только в междисциплинарных командах.

Наша исследовательская группа ставит задачу развития таких средств, фокусируясь на разработке методики синхронного и диахронного полуавтоматизированного анализа освещения острых социальных проблем в блогах и социальных сетях. В задачи входит обнаружение и прогнозирование социальных расколов и напряжений. В качестве тестовой темы взята тема ислама. Данный доклад посвящен рассказу о том, что сделано группой к настоящему времени в области семантического анализа: разработано ПО для загрузки, хранения текстов блогов и построения выборок, изучены алгоритмы и программные средства кластерного анализа и других способов классификации текстов, проведен пробный анализ на тестовых выборках. Предварительные результаты показывают общее тематическое членение краткосрочных выборок, долю и состав текстов об исламе. В дальнейшем планируется

расширение выборок, проведение анализа эмоциональной нагруженности различных тематических кластеров и анализ динамики освещения тестовой темы.

Social problems and change detection through a large-scale text analysis of blogs and social networks

For almost a decade social reality has got a new home – Internet that has become an indicator, a predictor and a promoter of social processes and change. “Arabic revolutions” of this year have demonstrated that it may play a crucial role even in societies with relatively low Internet penetration. Seemingly, Internet gives an unprecedented opportunity for sociologists to study and forecast social processes. However, academic community (not only that of sociologists) faces a necessity to master analysis of large amounts of text data, which is impossible without special software, advanced mathematical algorithms and computer linguistic approaches. Such sociological research has proven to be possible only in interdisciplinary teams.

The main goal of our research group is to develop such methods focusing on synchronous and diachronic semi-automatic analysis of coverage of social problems in blogs and social networks. Among them are detecting and forecasting social cleavages and tensions. The topic of Islam is taken as a test one. This paper addresses what has been already done by the group in the field of semantic analysis: a software has been developed allowing blog data download, storage and sample formation, algorithms of text clusterisation and classification have been studied, initial analysis of test samples has been carried out. Preliminary results show the general thematic distribution in short-terms samples, share and composition of texts about Islam in the sample. The next step will be widening of samples, sentiment analysis of thematic clusters and analysis of dynamics of coverage of the test topic.