DEVELOPING METHODS FOR SEMANTIC & NETWORK ANALYSIS OF BLOGS

Project leader: Olessia Koltsova, dean, department of sociology, Higher School of Economics (St.Petersburg branch)
Participants: Anastasia Kincharova, Tatiana Yefimova, Elisaveta Tereschenko, Yulia Pavlova (HSE St.Petersburg), Lidia Pivovarova (St.Petersburg State University)

Project supported by the grant of Research Fund of HSE (Moscow), 2011-2013

Research agenda

The general mission of this project is to develop methods of internet data analysis for sociological goals. A vast number of mathematical models, methodological approaches and particular methods, as well as algorithms and software aimed at internet analysis does exist in the domains of computer science, computational linguistics, and network analysis. However, they mostly do not address typical sociological tasks (apart from social network analysis which is however mostly non-internet-oriented). Some of those methods and software are commercial and do not reveal their algorithms; others still focus on practical tasks (e.g. data mining to detect online terrorist communities or marketing opinion mining); still others are aimed at mathematical modeling, software development or, e.g., taxonomies or thesauri building – all these can be seen only as tools for further sociological analysis.

Communities of social scientists (in particular, of sociologists and of media scholars) are largely unfamiliar with the listed above areas of knowledge. They import traditional research instruments and research logic into the completely new research domain that prevent them from studying Internet as a new social phenomenon. Sociologists mostly see it as a useful tool to assist them to work with traditional units of analysis – human beings who now can be polled online. At best, traditional statistical procedures are applied to Internet consumption, but not to its contents – it is here where sociological methods do not apply. Media scholars tend to see Internet as just another communication medium, a mixture of mail, newspaper and some others, while it is already a barometer of the social processes, and wider – a new, electronic "home" of social reality where those processes emerge and develop intertwining with off-line phenomena. Simultaneously, as a set of electronic signals, Internet is a huge repository and archive of social life, a database available of automatic analysis. But this hugeness is precisely the thing social scientists can not cope with.

Research goals

The goal of the project is to find ways to identify how a given socially significant topic is interpreted by an Internet-active part of society – in this case, by bloggers. Only Russian-language segment of the blogosphere is analyzed, and certain time limits will be set to it. As a topic exemplifying a hot social issue, we have selected Islam. Islam is the second largest religion in Russia, some of its adherents being immigrants, but others – indigenous peoples conquered by the Russian Empire back in 16-19 centuries. Tensions produced by these and other circumstances are reflected in the Russian blogosphere. The project seeks to detect what discourses (subtopics & opinions) on Islam exist in the blogosphere, which of them dominate, how they are (dis)connected to each other and if they correlate to the Internet communities.

The immediate tasks of the research, therefore, are:

- to extract a population (and samples) of all texts of a given topic (Islam) (with hyperlinks);

- to divide the corpus of texts into semantically similar clusters;
- to detect hyperlink-based subgraphs in the given corpus (community detection);
- to juxtapose semantic clusters and communities;
- to draw sociological conclusions about discourses and communities structures, sizes, and relations.

Each task presents a set of methodological and technical problems to be resolved.

Data and data processing

An agreement is signed with the leading Russian search engine Yandex about obtaining two entire one-month collections of blogposts and comments with metadata about bloggers and links, and a friending-based graph of the Russian blogosphere. The collections include blogs indexed by Yandex. Each one-month collection contains about 210 mln posts, including about 30 mln "proper" posts (i.e. not from micro-blogs), and about four times more comments. The files will be either xml or txt or MS SQL. Most probably an MS SQL server will then be developed to accommodate the collections. Relevant data of smaller sizes will be extracted by the means of SQL statements and converted into other formats to be imported to necessary software for network and text analysis.

Methodological problems to be resolved

Forming population. Sociologists usually do not have to "form" a population: it is "out there", usually unreachable, but easily discernable. Extracting a population of all texts dealing with a topic is a separate problem that may not be resolved without human coding and expertise. To make the procedure less dependent on certain individuals, such methods as intercoder reliability control and iterativeness up to saturation are most likely to be applied.

Sampling. If the population is too large for the existing software, an additional problem of forming samples will arise. Random sampling of texts from the database is very easy (unlike random sampling of people), but this kind of sample is not suitable for network analysis. However, according to preliminary assessments, sampling will not be needed.

Choosing unit of analysis. Being multi-topical, entire blogs can hardly be clusterized except by fuzzy clustering, but even then clusters will be blurred and distorted by too much noise. Freed from irrelevant posts blogs may produce less noisy clusters, still fuzzy clustering will be needed. To the present knowledge there is no software that can simultaneously handle amounts of data as large as is needed AND do fuzzy clustering. Since most posts are uni-topical, choosing a post as a unit of analysis may produce less noisy and less distorted clusters, but these results will be hardly comparable with SNA results. Intext links between posts are relatively rare, so reliance on them will produce very loose networks. Friending links connect blogs, not posts (in Russia blogplatforms and social media platforms are fused). Building commenting networks would be interesting as they could reveal real conversational subnetworks, but adding comments as nodes will produce isolated star networks since each comment is attached to only one post, and posts are not attached to each other in the way comments are. Combination of posts and comments as vertices, and of comment-links and intext links as edges will produce a complex and hardly interpretable graph. It will have to be represented as a set of graphs to be analyzed separately.

Text clustering. Apart from the standard problem of multi-dimensionality (which can be resolved in a way similar to that mentioned in population forming section), there is a specific problem of shortness of studied texts. Lexical composition may have no discriminative power. As an emergency scenario clusterization by tags may be applied. However, preliminarily we can expect to have at least some clear semantic clusters with a vast number of posts classified as noise. A separate problem is the software that simultaneously: a) can work with Russian (preprocessing etc); b) can handle large amounts of data; c) has transparent information about its algorithms; d) works with unpredictable number of clusters. (B) and (d) are usually incompatible since large amounts of data usually demand greedy algorithms based on partitional clustering. If no such software is found, a combination should be used, plus some demands (and tasks) should be reduced. Preliminary analysis of almost 40 software shows that further choice is limited to 4-5 products.

Network analysis. An abundance of traditional SNA software uses standard clustering techniques aimed at detection of small-size cohesive subgroups (such as cliques and their derivatives) which make sense only for relatively small and dense networks. Community detection in large sparse graphs is a new sphere that is experiencing a burst of algorithm creation. However, most of them are not only undeveloped into package software, but mostly not yet properly tested. Furthermore, there is no yet consensus about criteria of quality of those algorithms. The most popular modularity-based approaches (relying on the notion of modularity introduced by Newman and Girvan) can be now called a “temporary standard”; they, along with some others, are united into an add-on software to R language, still there is a question if R and its applications can handle the data of the size needed.

Perspective plans

Perspective goals of this research stretch beyond the current project. In future the methodology may be further developed to include monitoring of the topic dynamics, as well as multilanguage comparisons.