III студенческая социологическая межвузовская конференция. СПб: НИУ-ВШЭ, 2011

Терещенко Елизавета Александровна
Санкт-Петербургский филиал НИУ-ВШЭ, 1 курс магистратуры, факультет социологии
eliza.ter@gmail.com

Научный руководитель: Кольцова Елена Юрьевна, доцент кафедры социологии, декан факультета социологии

Elizaveta Tereschenko

"Sociological analysis of Russian-language blogs: selection of clusterization algorithms"
Large size of Internet texts as a new kind of sociological data presents a challenge for social scientists as it demands new, computerized methods of text analysis. The first part of this paper describes most typical approaches to automatic text clustering, as well as specific difficulties of clusterization of blogs (short texts clustering problem, analysis of texts with specific vocabulary). The second part covers steps of software & algorithm selection. It describes what criteria should be used for estimation and selection of software for clustering texts, and preliminary results.

Тезисы доклада «Социологический анализ русскоязычных блогов: выбор алгоритма кластеризации текстов»

Вопросы, которые освещаются в докладе, поставлены и разрабатываются в рамках проекта «Разработка методологии сетевого и семантического анализа блогов для социологических задач», поддержанного Научным Фондом ГУ-ВШЭ в рамках конкурса «Учитель-Ученики 2011-2012 гг.» Целью проекта является разработка комплексной методики решения типичных социологических задач, направленных на изучение социальной реальности, представленной в Интернете. В проекте изучается, каким образом в блогах отражается различная тематика (например, тематика Ислама) и как на основании этой тематики блоги можно классифицировать. В таком случае объектом анализа выступают посты блогов, которые есть не что иное, как массив текстов. При этом значительный размер этого массива не позволяет разделить тексты на группы вручную, поэтому исследователь стоит перед необходимостью пользоваться автоматизированными алгоритмами, воплощенными в специалиальном программном обеспечении (ПО) и неразрывно с ним связаннами.

Для такого тематического анализа текстов и последующей классификации блогов можно использовать различные методы кластеризации. Таким образом, общая задача проекта может быть сформулирована как осуществление тематической классификации текстов блогов, наша непосредственная задача – поиск алгоритма кластеризации, адекватного объекту анализа.

Кластеризация – это процесс разделения совокупности объектов на однородные группы (кластеры или классы). При этом группировка происходит таким образом, чтобы сходство между объектами было максимальным, если они принадлежат к одной группе, и минимальным – если к разным.

Для изучения больших объемов данных используют неиерархические методы кластеризации, основанные на разделении. К таким алгоритмам относят самый популярный алгоритм k-means; общим недостатком алгоритмов этого класса является необходимость предварительно указывать количество кластеров. Также важно учитывать, что один текстовый объект может относиться к нескольким кластерам одновременно (этот подход к кластеризации называется «нечеткая кластеризация»), на больших массивах данных такой подход пока почти не реализован.

Сходство текстов обычно определяется по наличию в них общих слов и по сходству частот этих слов; более сложные алгоритмы, работающие со словосочетаниями, на больших массивах пока не отработаны. Для всех методик этого типа коллекция, в которой находится n документов и m различных терминов, представляется в виде матрицы mxn, элементы которой – частоты терминов в соответствующих текстах. Т.о., каждый текст – вектор в m-мерном пространстве, поэтому такая форма представления текста называется векторной. Для приведения текста к векторной форме все слова в нем сначала должны быть приведены к своей основе – лемме.

В ходе анализа блогов был выявлен ряд их особенностей как текстов:

1) Большое количество блогов содержит короткие тексты, т.е. тексты с низкими частотами ключевых слов, что затрудняет кластеризацию документов на основе частотного анализа;

2) Блоги часто используют неформальную лексику, что затрудняет приведение слов к их леммам и правильный подсчет их частот.

Проверять влияние этих ограничений на результаты еще предстоит. Как уже говорилось, для анализа больших объемов текстов социолог вынужден пользоваться ПО, в основе которых лежат специализированные алгоритмы, и подбор ПО представляет собой отдельную, не только техническую, но и содержательную задачу.

Наиболее полный обзор алгоритмов кластеризации текстов сделан в статье [Andrews & Fox, 2007]. Для поиска необходимого для работы ПО были также использованы обзоры программ [Ландэ 2005] и [Carpineto et al.. 2009].

Для выбора ПО были разработаны признаки-характеристики, которым должны соответствовать программы. На этапе предварительного выбора ПО оценивалось по таким характеристикам, как доступность; возможность его работы с русским языком; возможность обрабатывать большие массивы данных. Для второго этапа важнейшей характеристикой являются используемые алгоритмы кластеризации, в т.ч такие их показатели как вычислительная сложность, быстродействие, способность самостоятельно определять количество кластеров и работать с многоразмерными коллекциями.

На первом этапе было выбрано следующее ПО: gCLUTO, HAMLET, TextAnalyst, Carrot2, PolyAnalyst и NeuroXL Clusterizer. В рамках второго этапа уже удалось протестировать первые три из указанных программ: была взята предварительно сформированная текстовая коллекция и на ней были опробованы алгоритмы кластеризации, предлагаемые этими программами.

Нами сделан краткий обзор протестированных программных продуктов.

1. gCLUTO [George Karypis, http://glaros.dtc.umn.edu/gkhome] – программа для кластеризации и визуализации результатов кластеризации. Плюсом использования данной программы является то, что используемые в ней алгоритмы опубликованы и известны их сильные и слабые стороны. Программа использует иерархический агломеративный и неиерархический методы кластеризации. Основная проблема том, что исследователю нужно самому, часто интуитивно, определять количество кластеров.

2.HAMLET [ESRC (National Centre for Research Methods) и Bruno Hopp GESIS (Leibniz-Institut für Sozialwissenschaften)] – программа, которая выполняет семантический и символический анализ текстов, может осуществлять иерархический и неиерархический кластерный анализ. Принципы работы алгоритмов данной программы нигде не опубликованы. При этом ограничения по объему входных данных в доступной тестовой версии затрудняют оценку работоспособности программы.

3. TextAnalyst [Microsystems, Ltd, http://www.analyst.ru] – программа с широким функционалом, включая реферирование и кластеризацию текстов. Максимальный объем анализируемой выборки текстов не ограничен, что является главным достоинством этой программы. Однако доступ к алгоритмам программы ограничен.

На втором этапе оценки программного обеспечения обзор шести отобранных программ будет закончен. Этот этап также будет включать апробацию программного

обеспечения, оценку его работы и качества кластеризации (которая может быть выполнена математическими методами или с привлечением экспертов).

Список использованных источников:

1. Ландэ Д.В. 2005. *Поиск знаний в Internet. Профессионакьная работа.* М.: Издательский дом "Вильямс".
2. Andrews, Nicholas O. and Edward A. Fox, Recent Developments in Document Clustering, October 16, 2007. Technical Report TR-07-35, Computer Science, Virginia Tech. http://eprints.cs.vt.edu/archive/00001000/
3. Carpineto C., Osiński S., Romano G., Weiss D. 2009. *A survey of Web clustering engines.* ACM Computing Surveys (CSUR). 3.
4. George Karypis, http://glaros.dtc.umn.edu/gkhome
5. ESRC (National Centre for Research Methods) и Bruno Hopp GESIS (Leibniz-Institut für Sozialwissenschaften)
6. Microsystems, Ltd, http://www.analyst.ru