Фамилия	Имя	Отчество	Факультет	Курс	Телефон	Контактный @mail
Никишина	Ирина	Александровна	НИУ ВШЭ НН, Факультет гуманитарных наук	4	89308038318	irina.nikishina@mail.ru

Ф.И.О. научного руководителя	Название доклада
Уткина И.Е.	«Автоматическое извлечение семантической информации о действующих лицах из новостных ресурсов»

Автоматическое извлечение семантической информации
о действующих лицах из новостных ресурсов

Автоматическая обработка естественного языка (Natural Language Processing) используется в различных областях науки. Одной из наиболее значимых областей применения NLP являются поисковые системы, используемые в настоящее время не только для глобального поиска в сети Интернет, но также и для поиска текстов определенной тематики для различных компаний, предприятий, государственных учрежденииий с последующим изучением действующих лиц каждой статьи.

В настоящее время известно достаточное небольшое количество научных публикаций, посвященных данной области. В большинстве случаев в исследованиях авторы описывают приложения, позволяющие автоматически извлекать из статей личные имена. В большинстве случаев, подобные приложения узконаправлены в своем функционировании и находятся в закрытом доступе, являясь собственостью компаний. Именно поэтому создание универсального поискового приложения, применимого для любой сферы, является одной из актуальных областей разработок.

Полагаем, что дальнейшие разработки в области автоматического анализа новостных лент могут быть использован в любой области, связанной с анализом текста и определением личных имен. Преимущество данного метода заключается также в использовании автоматического метода определения действующих лиц, а также теории графов и репрезентации взаимоотношений между персонажами.

Автоматический анализ действующих лиц в новостных ресурсах:

- позволяет автоматически определять список имен, содержащихся в тексте
- позволяет формировать кластеры, сооответсвующие персонажам (в каждый кластер входят всевозможные варианты имени, содержащиеся в тексте)
- позволяет строить семантические графы, отображающие семантические отношения между персонажами
- позволяет изучить социальные отношения персонажей (свойства социального графа)

Практической значимостью исследования является создание приложения, позволяющего анализировать новостные статьи любой тематики. Компьютерная программа также позволяет наглядно представить взаимное расположение и соотнесение персонажей, автоматически определяемых в ходе анализа статьи.

Для анализа работы программы были отобраны англоязычные и русскоязычные политические статьи с сайтов *nytimes.com* и *ria.ru* соответственно. Данные новостные ресурсы были выбраны в соответствии со следующими критериями:

- доступность и возможность осуществления автоматического извлечения текстов статей
- разделение статей по темам и наличие политических статей
- постоянное обновление контента веб-сайтов
- наличие достаточного количества материала для анализа

Политическая тематика статей также была отобрана неслучайно. Именно политические тексты характеризуются наличием личных имен, связанных между собой определенными отношениями, раскрываемыми в статье. Особую важность представляют собой тексты, посвященные событиям в области внешней политики, так как они позволяют не только рассматривать отношения между лицами одного государства, но и международные связи.

Семантический анализ содержания высказываний текста производится при помощи семантических сетей, в которых вершинами являются персонажи (для каждого персонажа также преставлена его характеристика), а ребра, соединяющие вершины между собой, являются предикатами, формирующими отношения действующих лиц между собой.

Полученные результаты и алгоритмы действия программы будут рассмотрены в ходе доклада.

Библиография

1. Codd, E.F. A Relational Model of Data for Large Shared Data Banks / E.F. Codd // Communications of the ACM, – MD Comput, 1970. – Vol.13 (6). – P. 377–387.
2. Manning, C. Information Extraction and Named Entity Recognition / C.D. Manning // Cambridge University Press. – 2012. – P.73
3. Last, M. Web Intelligence and Security: Advances in Data and Text Mining Techniques for Detecting and Preventing Terrorist Activities on the Web / M. Last, A. Kandel // IOS Press, 2010. – pp.142-14.
4. Pouliquen, B. Building and Displaying Name Relations using Automatic Unsupervised Analysis of Newspaper Articles / B. Pouliquen, R. Steinberger, C. Ignat, T. Oellinger // JADT2006 : 8es Journ´ees internationales d'Analyse statistique des Donn´ees Textuelles. – 2006. – 12 pp.
5. Roussopoulos N.D. A semantic network model of data bases. — TR №104, Department of Computer Science, University of Toronto, 1976.