

Egozoom -- сервис для исследования эгосетей

© А. Г. Гомзин © С. А. Ипатов © А. В. Коршунов
Институт Системного программирования РАН,
Москва

gomzin@ispras.ru ipatov@ispras.ru korshunov@ispras.ru

Аннотация

В настоящее время большой популярностью среди пользователей сети Интернет пользуются социальные сети. Анализ социальных сетей широко используется в ряде приложений и дисциплин. Особый интерес представляет анализ эгосетей, т.е. ближайших соседей отдельных пользователей. В данной статье представлен краткий обзор сервиса для исследования эгосетей социальной сети Twitter.

1 Описание сервиса

Сервис Egozoom (<http://egozoom.at.ispras.ru/>) предназначен для анализа так называемой эго-сети пользователя Twitter.

Подобные сервисы известны для социальных сетей Facebook (<http://touchgraph.com/facebook>) и Vkontakte (<http://yasiv.com/vk>):

Под эго-сетью понимается подграф социального графа, состоящий из центрального пользователя (пользователь сервиса), его соседей и всех ребер между участниками этой сети.

Описываемый сервис позволяет выполнять следующие виды анализа эго-сети:

- Исследование кластерной структуры эго-сети
- Определение демографических атрибутов
- Определение частых слов и хештегов
- Определение тональности текста сообщения

2 Методы анализа

2.1 Исследование кластерной структуры эго-сети

Характерным свойством социальных сетей является существование сообществ или кластеров пользователей, связанных между собой сильнее, чем с остальными пользователями сети. Egozoom позволяет автоматически выявлять такие сообщества пользователей. Каждый пользователь,

как правило, входит в несколько сообществ. Алгоритм поиска сообществ должен это учитывать.

В нашем приложении используется алгоритм SLPA[1], который позволяет находить сообщества пользователей с пересекающимися множествами вершин.

2.2 Определение демографических атрибутов

Отличительной особенностью социальной сети Twitter является отсутствие доступных данных о пользователе, таких как пол, возраст, статус отношений, уровень образования, отношение к религии и политические взгляды.

Egozoom использует методы интеллектуального анализа текста для определения демографических атрибутов пользователя. На текущий момент в сервисе поддерживается определение следующих атрибутов: пол, возраст, статус отношений, религиозные и политические взгляды, отношение к алкоголю и курению.

Для определения таких демографических атрибутов как пол, статус отношений, религиозные и политические взгляды используется алгоритм классификации на основе SVM.

Возраст, отношение к алкоголю и курению определяются с помощью алгоритма регрессии и отображаются в виде численных значений.

В качестве признаков используются символьные и словесные n-граммы [2].

2.3 Определение частых слов и хештегов

Интересы пользователя извлекаются из его сообщений. В Egozoom интересы представляются в виде распределений слов и хештегов, полученных из текстов сообщений пользователей.

В приложении для каждого пользователя отображаются слова и хештеги, наиболее часто встречающиеся в сообщениях пользователя.

2.4 Определение тональности текста сообщения

В своих сообщениях пользователь может выражать позитивное или негативное отношение к чему-либо. В Egozoom для каждого сообщения автоматически определяется его тональность (эмоциональная окраска). При этом используется

метод, основанный на использовании классификатора MaxEnt[5], в качестве признаков словесные и символьные n-граммы, а также результаты тематического моделирования.

3 Описание приложения

Приложение позволяет загрузить и проанализировать данные текущего пользователя. Для этого требуется авторизация. Кроме того, для демонстрации возможностей приложения доступны результаты анализа нескольких демо-аккаунтов. Для доступа к данным используется Twitter API.

3.1 Отображение графа, сообществ и демографических атрибутов

Для отображения интерактивного графа эго-сети используется Force-directed layout[4].

Выбранный пользователь выделяется на графе с помощью подсветки входящих и исходящих ребер: входящие ребра подсвечиваются зеленым цветом, исходящие – красным.

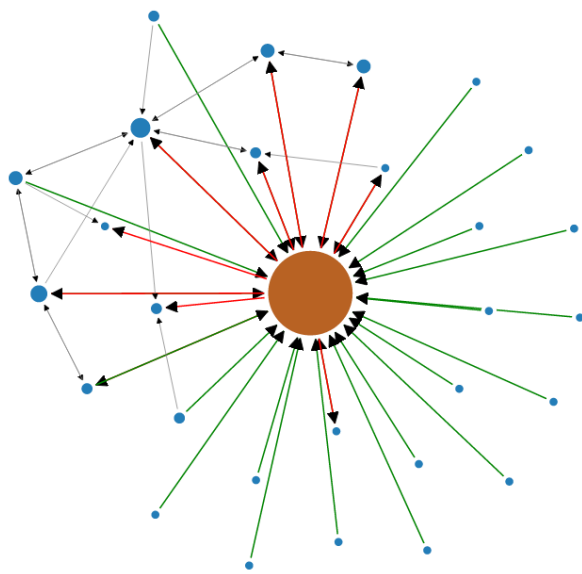


Рис.1 Эго-сеть пользователя Twitter

Сообщества на графе выделяются цветом вершин. Пользователи, входящие в одинаковые сообщества, подсвечиваются одним цветом. Некоторые пользователи могут входить сразу в несколько сообществ. В таком случае соответствующая вершина имеет несколько цветовых меток.

Для облегчения анализа сервис предоставляет следующие способы отображения демографических атрибутов на графе:

Фильтрация – позволяет оставить в отображаемом графе только вершины с заданным набором демографических атрибутов и/или типом связи. Вершины, которые не соответствуют заданному набору, удаляются из графа.

Подсветка – позволяет визуально разделить вершины с заданными набором значений демографических атрибутов. При этом, те вершины,

которые не соответствуют заданному набору атрибутов, не удаляются, а затеняются.

Распределения демографических атрибутов среди пользователей выбранного подграфа отображаются в правой части экрана в виде круговых диаграмм и гистограмм. При изменении отображаемого подмножества графа с помощью функции фильтрации, распределения автоматически пересчитываются. Это позволяет выявлять зависимости значений одних демографических атрибутов от значений других атрибутов.

3.2 Отображение сообщений выбранного пользователя

В зависимости от указанных настроек, приложения отображает сообщения пользователя за последнюю неделю или месяц с момента последнего сообщения. Также реализован поиск сообщений по хештегам.

Литература

- [1] Xie Jierui, Szymanski Boleslaw K, Liu Xiaoming. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process // 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW) / IEEE. 2011. С. 344–349
- [2] Anton Korshunov, Ivan Beloborodov, Andrey Gomzin, Christina Chuprina, Nikita Astrakhantsev, Yaroslav Nedumod, Denis Turdakov. Detection of demographic attributes of microblog users. // Proceedings of the Institute for System Programming Volume 25. 2013 y. pp. 179-194.
- [3] Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. Sentiment strength detection in short informal text. // Journal of the American Society for Information Science and Technology, Vol., 2544–2558. 2010.
- [4] Tutte, W. T. "How to draw a graph" // Proceedings of the London Mathematical Society 13 (52). 1963. 743–768.
- [5] Adam L. Berger t. Columbia University. Vincent J. Della Pietra, A maximum entropy approach to natural language processing // Computational Linguistics 22:1, March 1996

Egozoom: service for social ego-network analysis

Andrey Gomzin, Stepan Ipatov

Social networks are very popular among Internet users. Social network analysis is widely used in a number of applications and disciplines. Ego-networks (i.e. nearest neighbors of particular users) analysis is of particular interest. This article provides an overview of a service for Twitter ego-networks analysis.