

Магистрант И. А. Литвинович  
Науч. рук. ст. преп. А. С. Наркевич  
(кафедра программной инженерии, БГТУ)

## АЛГОРИТМЫ И ПОДХОДЫ К ОПРЕДЕЛЕНИЮ ПРИНАДЛЕЖНОСТИ ПРОФИЛЯ В СОЦИАЛЬНОЙ СЕТИ

Работа представляет собой анализ алгоритмов, используемых при написании магистерской диссертации, направленной на изучение и разработку алгоритмов поиска профилей в всемирно известных социальных сетях. Тема поиска профилей социальных сетей по фотографии является актуальной на сегодняшний день, так как социальные сети являются неотъемлемой частью жизни современного общества и одним из основных источников информации.

Пользователи социальных сетей выкладывают на своих страницах сотни фотографий, на которых фигурируют не только они сами, но и их друзья, знакомые, случайные люди, попавшие в кадр, животные и так далее. Для работы алгоритмов распознавания лиц, достаточно всего одной фотографии, например, аватарки. Но будет ли эта аватарка содержать фото профиля? Пользователи ставят на аватарки фотографии звёзд, а иногда профиль содержит только групповые фотографии. Таким образом, можно сделать вывод, что для определения принадлежности того или иного профиля конкретному человеку, встречающемуся на фотографиях этого профиля, необходимо разработать алгоритм и определить список наиболее важных признаков, по которым можно сделать однозначный вывод о принадлежности профиля.

Если рассматривать профили, которые состоят только из групповых фотографий, то можно определить владельца аккаунта, если учитывать его пол и возраст, а также друзей, профили которых были построены ранее.

Построение пользовательских профилей делится на следующие этапы.

- Выбор наиболее качественных фотографий пользователя. Если фотографий слишком много, необходимо использовать не более ста лучших. Качество фотографий определяется на основе:
  - наличия отметок пользователей на фото (фотопинов) ручным способом;
  - метаинформации фотографии (фото загружено с мобильного телефона, снято на фронтальную камеру, в отпуске);
  - фото было на аватарке.

- Поиск лиц на выбранных фотографиях.
- Вычисление характеристического вектора лица.

Кластеризация векторов. Задача этой кластеризации – определить, какой именно набор векторов принадлежит владельцу аккаунта. Основная проблема – это наличие друзей и родственников на фотографиях. Для кластеризации используется алгоритм DBScan.

Алгоритм DBSCAN был предложен Мартином Эстер, Гансом-Питером Кригель и коллегами в 1996 году как решение проблемы разбиения (изначально пространственных) данных на кластеры произвольной формы. Большинство алгоритмов, производящих плоское разбиение, создают кластеры по форме близкие к сферическим, так как минимизируют расстояние документов до центра кластера. Авторы DBSCAN экспериментально показали, что их алгоритм способен распознать кластеры различной формы.

Идея, положенная в основу алгоритма, заключается в том, что внутри каждого кластера наблюдается типичная плотность точек (объектов), которая заметно выше, чем плотность снаружи кластера, а также плотность в областях с шумом ниже плотности любого из кластеров. Ещё точнее, что для каждой точки кластера её соседство заданного радиуса должно содержать не менее некоторого числа точек, это число точек задаётся пороговым значением.

Определение лидирующего кластера.

Для каждого кластера необходимо рассчитать вес на основании:

- размера кластера;
- качества фотографий, по которым построены эмбеддинги в кластере;
- наличия фотопиков, привязанных к лицам из кластера;
- соответствия пола и возраста лиц в кластере с информацией из профиля;
- близость центроида кластера к профилям друзей, вычисленным ранее.

Коэффициенты параметров, участвующих в вычислении веса кластера, обучаются линейной регрессией.

Чтобы кластер считался лидером, нужно чтобы его вес был больше ближайшего конкурента на константу, рассчитанную на обучающей выборке. Если лидер не найден, необходимо еще раз перейти к пункту 2, но использовать большее число фотографий. Для некоторых пользователей необходимо сохранить два кластера. Такое бывает для совместных профилей — некоторые семьи имеют общий профиль.

Получение эмбеддингов пользователя по его кластерам.

Построение вектора, который будет характеризовать внешность

Секция информационных технологий  
владельца аккаунта, – «эмбеддинг пользователя».

Эмбеддинг пользователя – это центроид отобранного для него (лидирующего) кластера. Строить центроиды можно множеством разных способов. После многочисленных экспериментов было принято решение использовать самый простой из них: усреднение входящих в кластер векторов. Как и кластеров, эмбеддингов у пользователя может быть несколько.

## ЛИТЕРАТУРА

1. Handbook of Face Recognition - New York, Dover Publications, September 4, 2001, 272p
2. Reliable Face Recognition Methods: System Design, Implementation and Evaluation/ Harry Wechsler, Springer; 2007 edition (November 2, 2006), 329p
3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
4. Duda, R.O., Hart, P.E., Stork, D.G. Pattern classification. Second edition. New York: Wiley, 2000. 680 p.
5. Гонсалес Р., Вудс М. Цифровая обработка изображений. М.: Техносфера, 2012. 1104 с. Gonzales, R., Woods, R. Digital image processing. Third edition. New Jersey: Pearson Education Inc., Prentice Hall, 2008. 954 p

UDK 004.72

PhD student Ghassan Jaber  
Supervise N. V. Patsei  
(Department of software engineering, BSTU)

## SCALABLE CONTENT ROUTING SYSTEM FOR ICN NETWORKS

Information-Centric networks (ICN) is a future architecture that aims to cache content objects in the network nodes or routers and allow the access to these contents from any location by ensuring in-network storage for caching contents, decoupling the content from the host address. The decoupling between publisher and subscriber removes the role of IP address, which works only as an identifier, locator and enabling multiparty communication through replication. Various projects like DONA, NDN, PUR-SUIT, SAIL, COMET, CONVERGENCE, MobilityFirst, PSIRP, CBCB, NetInf, KBN presented proposals in the ICN field in order to suit the new