Идентификация пользователей социальных сетей в Интернет на основе социальных связей

Сергей Бартунов * Антон Коршунов †

Аннотация

В настоящее время мы переживаем бум социальных интернет-сервисов. Каждый год появляется множество как общенаправленных, так и нишевых социальных сервисов, и для активных пользователей Интернет типично иметь несколько профилей в различных социальных сетях. Обнаружение профилей, принадлежащих одному человеку, в нескольких социальных сетях, позволяет получить более полный социальный граф, что может быть полезно во многих задачах, таких как информационный поиск, интернет-реклама, рекомендательные системы и т.д. В данной работе предлагается оригинальная «JLA-модель» идентификации пользователей, основанная на модели условных случайных полей и совместно использующая как аттрибуты пользовательских профилей, так и социальные связи. Предложенный подход особенно полезен в случаях, когда информация о пользовательских профилях малополезна, недоступна или скрыта из соображений приватности. Эксперименты на данных из двух популярных в настоящий момент социальных сетей «Facebook» и «Twitter» показали, что данный подход работает эффективнее существующих решений и способен сопоставить профили, которые невозможно сопоставить, используя только информацию об аттрибутах.

*sbartunov@gmail.com, Институт системного программирования Российской академии наук, Россия, 109004, г. Москва, ул. А. Солженицына, дом 25.

†korshunov@ispras.ru, Институт системного программирования Российской академии наук, Россия, 109004, г. Москва, ул. А. Солженицына, дом 25.

Ключевые слова: идентификация пользователей, анализ социальных сетей, условные случайные поля, графические модели, обработка графов, машинное обучение

Введение

Еще несколько лет назад было трудно предположить, каким огромным будет присутствие социальных приложений в нашей жизни. Тем не менее, сейчас мы живем в эпоху онлайновых социальных сетей. Ввиду беспрецедентного масштаба социальных сервисов и, как следствие, большого количества информации, заключенной в них, привлечение социальной составляющей при решении многих задач может значительно улучшить результаты.

Основной проблемой при задействовании социальной информации является её фрагментированность среди множества различных онлайновых социальных сетей. Несмотря на то, что существуют попытки по обеспечению единого способа взаимодействия между различными социальными платформами (например, Google OpenSocial [1]), они не получили широкого использования, а новые социальные сервисы продолжают появляться. Процесс идентификации пользователей необходим для объединения различных социальных сетей и получения более полной картины о социальном поведении данного пользователя в «Интернет».

В данной работе мы фокусируемся на задаче идентификации пользователей в т.н. *локальной перспективе*. Это подразумевает сопоставление профилей в рамках списка контактов некоторого центрального пользователя. Такая задача часто возникает при работе с контактами в социальных мета-сервисах, которые, в частности, могут служить для объединения новостных потоков в поддерживаемых социальных сервисах (таких как «Path»[2]) или предоставления единой системы обмена сообщениями (сервисы «Meebo» и «imo»[3]). Другая область, в которой возникает подобная задача, это функция автоматического объединения контактов, часто присутствующая в современных мобильных устройствах (например, в смартфонах на платформе «Android»).

[1] http://code.google.com/apis/opensocial/
[2] http://path.com/
[3] http://imo.im/

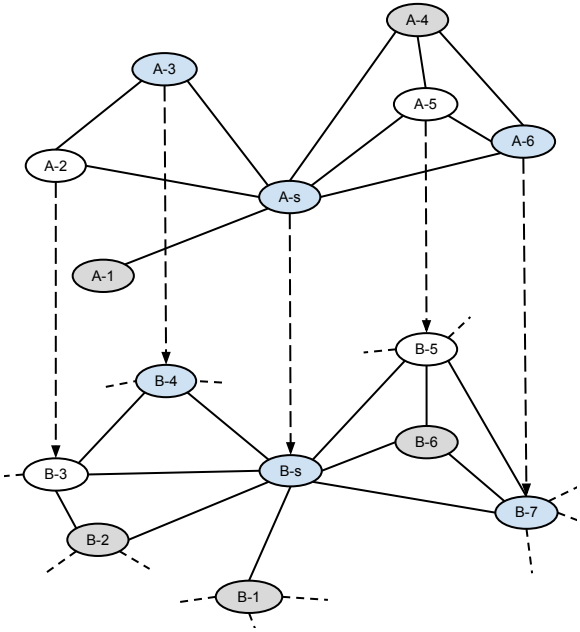

Рис. 1: Результат идентификации пользователей. Пунктирные стрелки обозначают проекции между профилями. Для вершин, закрашенных синим, проекции были известны заранее, проекции для незакрашенных вершин были установлены алгоритмом, для вершин, закрашенных серым, проекции не были найдены

Постановка задачи

Рассмотрим два социальных графа $< A, B >$. Под социальным графом будем понимать граф, узлы которого представлены пользовательскими профилями с различными аттрибутами (например, имя, день рождения, родной город и т.д.), а ребра социальными связями между профилями. Эти связи могут быть как направленными, так и ненаправленными в зависимости от семантики отношений, которые они представляют.

Задача идентификации пользователей заключается в поиске как можно большего числа правильно определенных пар профилей (v, u) таких, что $v \in A, u \in B$, принадлежащих одному и тому же реальному человеку. Сопоставленный профиль для профиля $v \in A$ мы

будем обозначать как $\mathrm{pr}(v) \in B$ и называть *проекцией* профиля $v \in A$ в B, а множество всех проекций $\{\mathrm{pr}(v)\}_{v \in A}$ профилей из A в B как $\mathrm{PR}(A)$. Если же для профиля $v \in A$ не найдено подходящей проекции, то проекцию для v будем называть *нейтральной* и обозначать как $\mathrm{pr}(v) = \mathbf{N}$. Пример двух таких социальных графов $< A, B >$ и сопоставленных пар профилей изображен на рисунке 1.

Так как в нашей работе мы рассматриваем задачу идентификации в *локальной перспективе*, то подразумевается, что графы A и B имеют структуру *эго-сетей* (англ. *ego-network*) некоторого пользователя. Эго-сеть вершины v представляет из себя граф, состоящий из вершины v и всех вершин, расстояние от которых до v не превышает двух. Такое ограничение отражает реальные ограничения использования социальных приложений, в которых для предоставлении какой-либо информации о социальных связях требуется непосредственное разрешение пользователя.

Обзор существующих методов

Идентификация пользователей

На момент написания данной работы наиболее значительным трудом по идентификации пользователей является работа Veldman [12], в которой представлено множество эвристик, использующих как информацию о профилях, так и связей между ними. Похожие исследования описаны в [8, 4, 9, 13]. Motoyama и др. [8] сопостовляли пары профилей между сетями «Facebook» и «MySpace», в свою очередь Gaewon и др. [4] решали аналогичную задачу для «Twitter» и «EntityCube». Raad и др. [9] в своем исследовании генерировали случайные социальные графы со случайно сформированными профилями и применяли к ним многочисленные сложные эвристики с целью не упустить ни одного потенциально полезного источника информации, доступного в социальной сети. В работе Vozecky и др. [13] профили из «Facebook» и «StudiVZ» представлялись как векторы признаков, к которым в последствии применялись операции точного, частичного и нечеткого сравнения, по результатам которых проводилась идентификация.

Общая схема работы систем идентификации пользователей

Несмотря на то, что описанные выше исследования применялись к данным самых разных социальных сервисов, не составляет труда выделить и проанализировать общую схему работы этих систем:

1. Приведение данных из полей профилей из двух социальных сетей к некоторому общему виду (например, вектору, элементами которого являются поля профилей)

2. Попарное применение техник нечеткого сравнения между профилями из одной сети и профилями из другой

3. Подсчет результирующего показателя *похожести* между профилями и отсечение всех парных результатов, для которых этот показатель ниже некоторого порогового значения

После этого все оставшиеся пары считаются сопоставленными между собой и принадлежащими одному пользователю.

Несмотря на относительно неплохое качество работы этих систем они все имеют общий недостаток - слишком простую модель сравнения текстовых аттрибутов профилей. При этом социальная информация не учитывается, либо учитывается слишком слабо. При этом информация, содержащяся в профилях, достаточно ненадежна, так как данные, указанные пользователям, в разных социальных сетях могут сильно отличаться, быть скрытыми из-за настроек приватности или не поддерживаться в актуальном состоянии.

Для улучшения этого общего подхода необходимо привлечь дополнительные источники данных, в частности информацию о социальных связях. В некоторых работах для этого применяется техника сравнения *частично сопоставленных* списков контактов [12], которая заключается в подсчете показателя похожести между множествами профилей, которые ранее были сопоставлены по именам. Очевидно, что подобная эвристика может привести к *смещению* в результатах. Подход, представленный в этой работе, активно использует социальные связи обеих рассматриваемых социальных сетей путем сравнения оригинальных списков контактов, естественным образом комбинируя их с информацией аттрибутов профилей, благодаря чему лишен многих недостатков существующих систем идентификации пользователей.

Разрешение сущностей

Помимо описанной выше задачи идентификации пользователей, существует также ряд близких задач, результаты которых могут быть использованы и в применении к объединению социальных графов. Одной из таких задач является *разрешение сущностей* (англ. *entity resolution*), которая заключается в определении записей базы данных, относящихся к одному и тому же объекту реального мира (не обязательно описывающие его). В работе [10] авторы строят сеть марковской случайной логики, узлами которой являются атомарные утверждения о записях базы данных с весом от 0 до 1 в зависимости от истинности или ложности утверждения, а ребрами логические связи между ними, после чего ищут оптимальную (*наиболее правдоподобную*) конфигурацию истинности утверждений, используя информацию о логических зависимостях. Подобный подход был также применен и в работе [11] для задачи устранения дубликатов (англ. record deduplication) в графе цитирования авторов научных статей, где вместо марковской случайной логики применены условные случайные поля [6].

Аналогичный подход, наиболее похожий на представленный в данной работе, описан в [11], где для устранения дубликатов (англ. record deduplication) в графе цитирования авторов научных статей используются условные случайные поля [6]. Основной идеей было построить условное случайное поле и представить узлами атомарные утверждения вида “являются ли эти две записи дубликатами?” с возможными значениями “да” или “нет” и узлы-улики с информацией о близости аттрибутов рассматриваемых объектов. Ребра же между утверждениями обозначали бы *условную зависимость* между ними, в то время как ребра между утверждениями и уликами - правдоподобность данного утверждения при известном значении похожести аттрибутов. После чего из данной модели возможно сделать *вывод* оптимальной конфигурации ответов на утверждения.

Две данные работы демонстрируют эффективность решения задачи, похожей на идентификацию пользователей как совокупности нескольких взаимозависимых задач, а также применения графических вероятностных моделей. Тем не менее, «JLA-модель» хоть и также основана на условных случайных полях, но значительно отличается от описанных выше подходов следующими аспектами:

- Более компактное и в то же время более естественное представление модели условных случайных полей, которая строится на

основе одного из социальных графов, а не графе связанных утверждений

- Помимо информации о строковой похожести аттрибутов объектов используется информация о графовой близости вершин

- Размер и подробность графической модели в описанных выше подходах делают вывод ответа неэффективным на относительно больших данных, в то время как предлагаемый в данной работе подход использует более компактное представление и при помощи описанных техник оптимизации делает возможным вывод решения даже для больших социальных графов, в том числе параллельно.

«JLA-модель»

«JLA-модель» (от англ. joint link-attribute), представленная в данной работе, основывается на следующих соображениях:

1. Необходимо совместно использовать как аттрибуты профилей, так и социальные связи между ними

2. Задачи выбора проекций для связанных вершин в графе A взаимосвязаны, иначе говоря, выбор проекции для некоторой вершины зависит от значений проекций связанных с ней вершин.

3. Если две вершины в графе A связаны, их проекции должны иметь как можно меньшие расстояния в графе B.

В данной работе из соображений простоты и эффективности в качестве функции расстояния в графе B используется коэффициент Дайса:

$$\mathsf{network\text{-}distance}(v,u) = 1 - \frac{2 \cdot w(L_v \cap L_u)}{w(L_v) + w(L_u)}, v, u \in B,$$

где L_v и L_u - множества вершин, связанных с v и u соответственно, а $w(L) = |L|$ - вес этих множеств. Причем $0 \leq \mathsf{network\text{-}distance} \leq 1$.

Предполагается, что один из графов $< A, B >$ является ненаправленным. В дальнейшем без ограничения общности будем считать таким графом A.

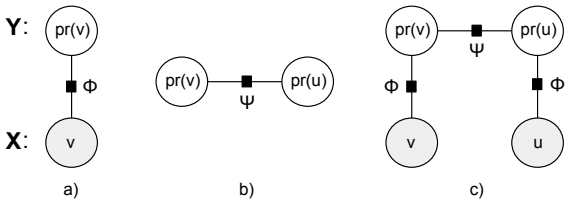

Рис. 2: Структура «JLA-модели». a) идентификация, основанная на аттрибутах профилей b) идентификация, основанная на социальных связях c) полная модель

На основе графа A строится модель условных случайных полей [6], в которой множество наблюдаемых переменных представлено вершинами графа A: $\mathbf{X} = \{\mathbf{x}_v \mid v \in A\}$, с каждой из которых ассоциирована одна скрытая переменная $\mathbf{Y} = \{\mathbf{y}_v \mid v \in A\}$, определяющая проекцию данной вершины: $\mathbf{y}_v = \mathrm{pr}(v) \in B$. Эти пары переменных связаны фактором унарной энергии Φ. Две скрытые переменные $\mathbf{y}_v$ и $\mathbf{y}_u$ связаны фактором бинарной энергии Φ тогда и только тогда, когда связаны вершины v и u в графе A.

Совместная природа модели выражается в том, что похожесть аттрибутов профилей учитывается при помощи унарной энергии Φ, а социальные связи - через бинарную энергию Ψ (см. рис. 2). Это делает модель адаптивной по отношению к данным, которые доступны для использования. Так, если отсутствует информация о социальных связях, то $\Phi \equiv 0$ и модель приобретает форму стандартной системы идентификации. В то же время, если данные анонимизированны, то есть, социальные связи присутствуют, но вся полезная для идентификации информация в профилях убрана, то $\Psi \equiv 0$ и используется только структура социальных графов.

Таким образом, модель порождает следующее вероятностное распределение:

$$p(\mathbf{Y}|\mathbf{X}) = \exp(-E(\mathbf{Y}|\mathbf{X})),$$
$$E(\mathbf{Y}|\mathbf{X}) = \sum_{v \in V} \Phi(\mathbf{y}_v|\mathbf{x}_v) + \sum_{(v,u) \in E} \Psi(\mathbf{y}_v, \mathbf{y}_u),$$

где E это *функционал энергии*, моделируемый функцией *унарной энергии* Φ и функцией *бинарной энергии* Ψ. Обе энергетические функции вещественны и неотрицательны.

Таблица 1: Схема сравнения полей профилей в сетях Facebook и Twitter

Facebook	**Twitter**	**Функция сравнения**
Name	Name	VMN
	Screen name	Screen Name measure
Website	URL	URL measure

Унарная энергия отвечает за похожесть профиля в A и его проекции в B с точки зрения полей профилей:

$$\Phi(\mathbf{y}_v|\mathbf{x}_v) = \alpha(v) \cdot \mathsf{profile\text{-}distance}(v, \mathrm{pr}(v)),$$

а бинарная энергия отвечает за близость между проециями вершин v и u в графе B:

$$\Psi(\mathbf{y}_v, \mathbf{y}_u) = \mathsf{network\text{-}distance}(\mathrm{pr}(v), \mathrm{pr}(u)).$$

Здесь $0 \leq \mathsf{profile\text{-}distance} \leq 1$, и $\alpha(v) = \log(degree(v)) \geq 0$ - коэффициент баланса между унарной и бинарной энергией.

Для двух данных графов $< A, B >$ существует оптимальная конфигурация проекций:

$$\mathbf{Y}^* = \underset{Y}{\mathrm{argmin}} E(\mathbf{Y}|\mathbf{X}), \tag{1}$$

которая минимизирует функционал энергии, максимизируя при этом правдоподобие модели.

Похожесть профилей

Для определения похожести профилей из разных социальных сетей необходимо учесть все доступные поля, содержащиеся в них, с использованием различных функций нечеткого сравнения. Для рассматриваемых в данной статье сетей «Twitter» и «Facebook» использовалась схема сравнения, изображенная в таблице 1.

«URL measure» проверяет, упоминается ли в одном профиле URL второго профиля. «Screen Name measure» проверяет на полное совпадения имени в «Facebook» и отображемого в адресе имени пользователя в «Twitter». VMN это функция близости, заимствованная из [13].

Таблица 2: Сравнение классифакторов похожести унарной энергии

алгоритм	полнота	точность	F_1
Naive Bayes	**0.862**	0.308	0.453
C4.5	0.569	0.86	0.685
C4.5 с MultiBoosting	0.669	**0.879**	**0.76**

Путем применения функций близости к соответсвующим полям двух профилей, формируется *вектор похожести* $V(v, \mathrm{pr}(v))$. Причем если хотя бы одном из профилей поле отсутствует или недоступно, то соответствующий элемент ветока V неопределен. Вектор V используется как набор признаков, на которых обучается специальный бинарный классификатор, определяющий принадлежат ли профили v и $\mathrm{pr}(v)$ одному и тому же человеку.

Таким образом, можно определить:
$\mathsf{profile\text{-}distance}(v, \mathrm{pr}(v)) = \mathrm{P}(\text{разные люди}|V(v, \mathrm{pr}(v))$, поскольку и унарная энергия, и вероятность, возвращаемая классификатором принадлежат интервалу $[0, 1]$. Сравнение различных алгоритмов классификации при помощи кросс-валиации с 3-я разбиениями представлено в таблице 2.

Алгоритм C4.5 с MultiBoosting был выбран для дальнейший экспериментов, как показавший наибольшую эффективность. Следует отметить, что ни один из классификаторов не смог „объяснить" принадлежность профилей на основании только полей профилей.

Заранее известные проекции

«JLA-модель», также как и многие другие алгоритмы, использует информацию о заранее известных проекциях (в зарубежной литературе *anchor nodes* или *seed nodes*). Такие проекции могут быть сообщены алгоритму перед началом работы, или просто могут считаться таковыми если $\mathsf{profile\text{-}distance}(v, \mathrm{pr}(v)) \leq \Delta$.

Для каждой вершины v с заранее известной проекцией $\mathrm{anchor}(v)$ значения энергий зафиксированы:

$$\begin{aligned} &\Phi(\mathbf{y}_v \mid \mathbf{x}_v) = \infty \\ &\Psi(\mathbf{y}_v, \mathbf{y}_u) = \Psi(\mathbf{y}_u, \mathbf{y}_v) = \infty \end{aligned} \qquad \text{if } \mathbf{y}_v \neq \mathrm{anchor}(v)\ \forall u$$

Заранее известные проекции повышают точность полученных результатов, а также значительно уменьшают вычислительное время работы алгоритма. Информация распространяется от вершин с

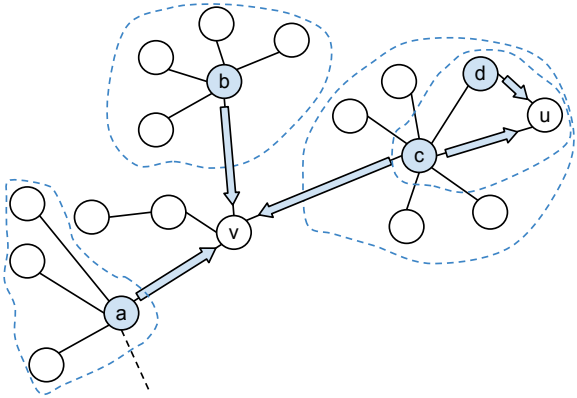

Рис. 3: Распространение информации от вершин с заранее известными проекциями. Проекции для вершин внутри областей, размеченных пунктирами, могут быть установлены независимо.

заранее известными проекциями (см. рис. 3), и для подграфов графа A, связанных с остальным графом только посредством таких вершин, проекции могут быть установлены независимо, что позволяет параллельно обрабатывать большие графы.

Нейтральные проекции и очистка результатов

Поскольку выбрать разумные фиксированные значения функций близости для нейтральных проекций не представляется возможным, для очистки результатов (1) от неправильно выбранных проекций необходимо привести соответствующую процедуру.

Очевидной методикой очистки результатов может служить повторный вывод ответа (1) при построении модели на графе B (то есть, с противоположным направлением проецирования) с последующим удалением всех вершин, для которых выбранная проекция не совпала. Иначе говоря, в результаты попадают только вершины, которые были *взаимно* размечены при обратном направлении проецирования. Несмотря на простоту и интуитивность, данная процедура очистки требует повторогого вывода ответа, что может быть слишком затратным с точки зрения времени выполнения для относительно больших графов, а также достаточно груба, так как не учитывает *причину*, по которой была допущена ошибка.

Таблица 3: Эффективность классификаторов очистки

алгоритм	полнота	точность	F_1
Naive Bayes	0.762	0.256	0.383
Support Vector Machine	0.662	0.935	0.775
C4.5	0.715	**0.939**	0.812
C4.5 с MultiBoosting	**0.844**	0.902	**0.872**

В качестве более продвинутого решения предлагается схема обучения бинарного классификатора (C4.5 с MultiBoosting), который, используя информацию о контекстке каждой вершины в A, решает, правильно ли для неё выбрана проекция. Для этого используются следующие признаки:

1. $\mathsf{profile\text{-}distance}(v, \mathrm{pr}(v))$
2. Средняя графовая близость к проекциям смежных вершин
3. Доля заранее известных проекций среди смежных вершин
4. Взаимо-согласованность смежных вершин с заранее известными проекциями:

$$\frac{1}{n} \cdot \sum_v \frac{1}{n-1} \sum_{u \neq v} \mathsf{network\text{-}distance}(\mathrm{pr}(v), \mathrm{pr}(u))$$

Сравнение различных алгоритмов классификации при кросс-валидации с 3-я разбиениями приведено в таблице 3.

Результаты работы

Предложенный в данной работе подход был протестирован на данных из двух наиболее популярных на данный момент социальных сетей «Facebook» и «Twitter». Для 16 центральных пар профилей в обоих социальных сетях были загружены и размечены „эго-сети", преимущественно самими владельцами профилей. Эти основные данные были использованы для настройки всех алгоритмов машинного обучения и для тестирования точности всех алгоритмов с использованием кросс-валидации с 3-я разбиениями.

Таблица 4: Экспериментальные данные

	Twitter	**Facebook**
Основная выборка		
# центральных пользователей	16	
# профилей	398	977
# связей	1 728	10 256
# сопоставленных профилей	141	
# заранее известных проекций	71	
Дополнительная выборка		
# центральных пользователей	17	
# профилей	1 499	7 425
# связей	15 943	172 219
# сопоставленных профилей	161	

Кроме того, для тестирования «JLA-модели» была составлена дополнительная тестовая выборка. Поскольку без привлечения владельцев аккаунтов социальных сервисов не возможно достоверно разметить данные, то дополнительная выборка использовалась в полуавтоматическом режиме для задачи *повторной идентификации*.

Полная статистика по использованным данным содержится в таблице 4.

Поскольку связи в социальной сети «Twitter» направленные и имеют семантику подписки, а не дружбы как в «Facebook», то при построении эго-сети было решено рассматривать только вершины, которые взаимно подписаны друг на друга, для симуляции отношений дружбы. Таким образом, при построении модели условных случайных полей на графе «Twitter» использовался ненаправленный граф, как того требует модель. При расчете расстояний на графе «Twitter» в качестве списка контактов также использовались списки взаимно подписанных друг на друга профилей.

Базовые алгоритмы

В качестве базовых алгоритмов, с которыми проводилось сравнение «JLA-модели», были выбраны алгоритмы, использующих только информацию о полях профилей, поскольку именно так работает

большинство систем идентификации пользователей. Базовые алгоритмы сопоставляют каждому профилю из графа A не более одного профиля из графа B, так чтобы с одной стороны максимизировалась некоторая функция близости между профилями и их проекциями, а с другой стороны не было двух и более профилей, спроецированный в один и тот же профиль в графе B. Таким образом, базовые алгоритмы решали задачу *оптимального парасочетания*.

Рассматриваемые алгоритмы использовали следующие функции близости:

1. Взвешенная сумма элементов вектора похожести $V(v, \mathrm{pr}(v))$. Веса подбирались при помощи линейной регрессии, исходя из предположения, что для правильно выбранных проекций, сумма должна равняться 1.

2. $\mathsf{profile\text{-}distance}(v, \mathrm{pr}(v))$. Иначе говоря, данный алгоритм использовал ту же функцию похожести, что и «JLA-модель».

Базовые алгоритмы также имели пороговые значения, ниже которых значения похожести не рассматривались. Эти значения были настроены таким образом, чтобы достигалась максимальная точность, так как именно это ожидается от реальной системы идентификации.

Оценка качества алгоритмов

Рассматриваемые алгоритмы оценивались с точки зрения общеизвестных метрик *точности* и *полноты*:

$$\text{полнота} = \frac{\text{true-positives}}{\text{true-positives} + \text{false-negatives}}$$

$$\text{точность} = \frac{\text{true-positives}}{\text{true-positives} + \text{false-positives}}$$

При тестировании «JLA-модель» использовалась по умолчанию с процедурой очистки результатов при помощи обученного классификатора, а также с наивной техникой взаимной проекции при противоположных направлениях. Для получения списка заранее известных проекций использовались результаты второго базового алгоритма.

Результаты оценок качества алгоритмов отображены в таблице 5. Практически все алгоритмы достигли максимальной точности.

Таблица 5: Оценка качества алгоритмов на основной выборке

алгоритм	полн.	точн.	F_1
безразличные к направлению проекции			
Базовый 1 (взвешенная сумма)	0.45	0.94	0.61
Базовый 2 (вероятностная похожесть)	0.51	**1.0**	0.69
JLA, взаимн. проекц., аноним.	0.6	**1.0**	0.76
JLA, взаимн. проекц.	0.66	0.99	0.79
Twitter → Facebook			
JLA, анонимн. ($\Phi \equiv 0$)	0.62	**1.0**	0.77
JLA	0.79	**1.0**	0.89
Facebook → Twitter			
JLA, анонимн. ($\Phi \equiv 0$)	0.61	**1.0**	0.76
JLA	**0.8**	**1.0**	**0.89**

Таким образом, основной трудностью было, сохраняя высокий показатель точности, сопоставить как можно большую часть профилей, тем самым достигнув максимального показателя полноты.

Второй базовый алгоритм незначительно обогнал первый алгоритм и обозначил предел возможностей систем идентификации, использующих только аттрибуты профилей.

Подход, использующий «JLA-модель», в среднем на 29% превзошел второй базовой алгоритм по показателю полноты, сохранив при этом максимальную точность. При этом включенная техника взаимного проецирования значительно занизила полноту, тем самым эмпирически подтвердив целесообразность использования классификатора для очистки результатов.

«JLA-модель» даже в условиях анонимизированных данных, но при наличии 50% заранее известных проекций, позволила дополнительно сопоставить около 10% процентов профилей, иначе говоря, *деанонимизировать* их. Это подтверждает как значимость социальных связей в задачах идентификации и деанонимизации пользователей, так и высокую для них пригодность «JLA-модели».

Повторная идентификация

Несмотря на то, что направление проекции на основной выборке практически не отразилось на результатах, некоторые отличия можно наблюдать при тестировании «JLA-модели» на дополни-

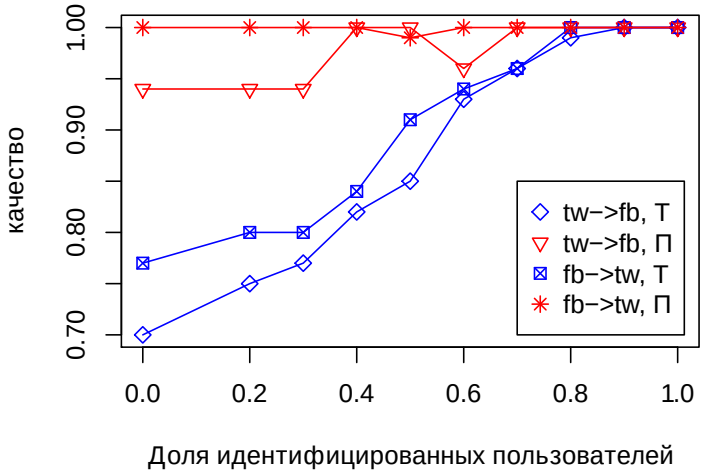

Рис. 4: Влияние доли известных идентифицированных пользователей на качество «JLA-модели»

тельной выборке.

Поскольку дополнительная выборка не была размечена вручную, на ней невозможно адекватно оценить качество работы алгоритма. Тем не менее, можно оценить, как хорошо алгоритм способен повторно идентифицировать профили, которые ранее были сопоставлены по профилям при помощи второго базового алгоритма.

Некоторая часть таких профилей фиксируется случайным образом, после чего у всех возможных проекций удаляется вся информация из профилей. Таким образом, правильные проекции для них могут быть найдены только благодаря связым и информации в виде оставшихся идентифицированных профилей.

На рисунке 4 отражена зависимость показателей точности и полноты среди идентифицированных профилей в зависимости от числа профилей с известными проекциями в среднем. Проекции, которые алгоритм определил для всех остальных профилей не учитываются. Поскольку социальный граф «Facebook» сильно более связный чем «Twitter» (согласно таблице 4), то при построении модели на графе «Facebook» даже при малой доле заранее известных проекций, информация от них распространялась лучше, и таким образом в среднем достигалось более высокое качество. Этот эксперимент демонстрирует значимость связности при выборе графа

для построения вероятностной модели, а также показывает, что при знании 80% проекций «JLA-модели» удалось найти остальные 20%.

Заключение

Результаты экспериментов на данных актуальных и в то же время небогатых по возможностям сравнения профилей социальных сетей показали важность социальных связей в задаче идентификации пользователей, а также их эффективное использование в предложенной «JLA-модели».

Несмотря на успешное применение модели для локальной перспективы, перенос её для глобальной перспективы нетривиален и не представляется возможным без использования достаточно большого числа заранее известных проекций. Это связано в первую очередь с большой вычислительной сложностью процесса вывода решения в условных случайных полях, который потребует значительных оптимизаций. Одна из таких оптимизаций - разбиение задачи на независимые подзадачи была предложена в данной статье, однако помимо этого также необходимо сужение множества возможных проекций для каждой вершины.

Открытым вопросом также является, как работает предложенный подход с социальными графами различных топологий, а не только с эго-сетями, а также насколько «JLA-модель» устойчива к потере или намеренному искажению информации о социальных связях.

Список литературы

[1] S. Bortoli, H. Stoermer, P. Bouquet (2007). *Foaf-O-Matic - Solving the Identity Problem in the FOAF Network.* In: Proceedings of the Fourth Italian Semantic Web Workshop (SWAP2007), Bari, Italy, Dec.18-20, 2007.

[2] P. Bouquet, S. Bortoli (2010). *Entity-centric Social Profile Integration.* In: Proceedings of the International Workshop on Linking of User Profiles and Applications in the Social Semantic Web (LUPAS 2010) 52-57.

[3] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens. *Random-walk computation of similarities between nodes of a graph, with application*

to collaborative recommendation. IEEE Transactions on Knowledge and Data Engineering, vol. 19, No. 3, March 2007.

[4] Gae-won Y., Seung-won H., Zaiqing N., Ji-Rong W. *SocialSearch: Enhancing Entity Search with Social Network Matching.* EDBT 2011.

[5] H. Kopcke, E. Rahm. *Frameworks for entity matching: A comparison.* Data & Knowledge Engineering, Vol. 69, No. 2. (2010), pp. 197-210.

[6] J. D. Lafferty, A. McCallum, P. McCallum, C. N. Fernando. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* Proceedings of the Eighteenth International Conference on Machine Learning, 2001.

[7] M. Lenzerini. *Data Integration: a Theoretical Perspective.* In PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems, pages 233-246. 2002.

[8] Motoyama, M., Varghese, G. *I Seek You - Searching and Matching Individuals In Social Networks.* WIDM '09: Proceeding of the eleventh international workshop on Web information and data management.

[9] Raad, E., Chbeir, R., Dipanda, A. *User Profile Matching in Social Networks.* 13th International Conference on Network-Based Information Systems (NBiS), 2010.

[10] P. Singla, P. Domingos. *Entity Resolution with Markov Logic.* In Proc. of the Sixth International Conference on Data Mining (ICDM'06).

[11] P. Singla, P. Domingos. *Multi-relational Record Linkage.* KDD Workshop on Multi-Relational Data Mining (pp. 31-48), 2004.

[12] Veldman, I. (2009) *Matching Profiles from Social Network Sites.* Master's thesis, University of Twente.

[13] Vosecky, J., Dan Hong, Shen, V.Y. *User identification across multiple social networks.* In Proc. of First International Conference on Networked Digital Technologies, 2009.