УДК 004.021:303.711:[311.214+303.722]

УЛУЧШЕНИЕ КАЧЕСТВА ИСХОДНЫХ ДАННЫХ В ЗАДАЧАХ МОДЕЛИРОВАНИЯ ИНТЕРНЕТ-СООБЩЕСТВ НА ОСНОВЕ КОМПЛЕКСНОГО ПРИМЕНЕНИЯ МОДЕЛЕЙ СЕГМЕНТАЦИИ, ИМПУТАЦИИ И ОБОГАЩЕНИЯ ДАННЫХ

О. О. Слабченко, В. Н. Сидоренко
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E–mail: vnsidorenko@gmail.com

Сделан обзор и анализ существующих подходов к улучшению качества данных на основе методов импутации и обогащения данных. Исследованы особенности структуры данных персональных аккаунтов социальных сетей и выявлены проблемы, отрицательно влияющие на их пригодность к анализу. Показано, что предварительная сегментация данных позволяет выявить их негомогенную структуру. Проведен эксперимент по восстановлению пропущенных значений посредством комбинации методов ассоциативных правил и Most common value на полном множестве данных и на отдельном сегменте. Установлено, что использование методов импутации применительно к каждому сегменту в отдельности позволяет повысить процент верно восстановленных пропущенных значений. Предложен подход к улучшению качества данных из социальных сетей путём комплексного применения моделей сегментации, импутации и обогащения данных.

Ключевые слова: комплексные сети, импутация пропущенных значений, обогащение данных, Social Mining, социально-сетевой анализ.

ПОКРАЩЕННЯ ЯКОСТІ ПЕРВИННИХ ДАНИХ В ЗАДАЧАХ МОДЕЛЮВАННЯ ІНТЕРНЕТ-СПІВТОВАРИСТВ НА ОСНОВІ КОМПЛЕКСНОГО ЗАСТОСУВАННЯ МОДЕЛЕЙ СЕГМЕНТАЦІЇ, ІМПУТАЦІЇ І ЗБАГАЧЕННЯ ДАНИХ

О. О. Слабченко, В. М. Сидоренко
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, г. Кременчук, 39600, Україна. E–mail: vnsidorenko@gmail.com

Виконано огляд і аналіз існуючих підходів до покращення якості даних на основі методів імпутації і збагачення даних. Досліджено особливості структури даних персональних акаунтів соціальних мереж і виявлено проблеми, які негативно впливають на їх придатність до аналізу. Показано, що попередня сегментація даних дозволяє виявити їх негомогенну структуру. Проведено порівняльний експеримент по відновленню пропущених значень шляхом комбінації методів асоціативних правил і Most common value на повній множині даних і на окремому сегменті. Експериментально встановлено, що використання методів імпутації стосовно кожного сегмента окремо дозволяє підвищити процент вірно відновлених пропущених значень. Запропоновано підхід до покращення якості даних із соціальних мереж шляхом комплексного застосування моделей сегментації, імпутації і збагачення даних.

Ключові слова: комплексні мережі, імпутація пропущених значень, збагачення даних, Social Mining, соціально-мережевий аналіз.

АКТУАЛЬНОСТЬ РАБОТЫ. Социально-сетевой анализ [1] (Social Network Analysis, SNA) приобретает большую популярность благодаря успешному развитию интернет-технологий, социальных сетей и электронной коммерции. Социальная сеть в интернете – один из видов комплексных сетей (complex networks), математических моделей феномена взаимодействия между явлениями, имеющими место в реальной жизни, и представляемых в виде графа с нетривиальными топологическими особенностями [2, 3]. Узлами такого графа являются социальные объекты, а связями – социальные взаимоотношения. Одним из главных объектов исследований SNA является характер связей между пользователями соцсетей и образование на их основе особых структур – онлайн-сообществ. С развитием сетевых коммуникаций такие сообщества становятся источником данных для анализа и выработки стратегии ведения бизнес-процесса, а также целью для распространения информации и влияния. Возможность сбора публично доступных данных о пользователях, их связях, взаимодействиях и наличие средств для этого позволяют решить широкий круг задач в сфере выявления и моделирования интернет-сообществ и информационного влияния [4]: формирование брендов компаний, распространение влияния, вирусный маркетинг, сегментация клиентов [5] и т.п.

Вышеперечисленные задачи, по сути, являются разновидностью задач Data Mining (DM) – Social Mining (SM). На практике при реализации DM-проекта неудовлетворительное качество данных становится одной из центральных проблем, так как для получения достоверных результатов исходная информация должна удовлетворять требованиям полноты, точности, достоверности и др. Поэтому применение методов, позволяющих улучшить качество исходной информации, необходимо ещё на этапе ETL-процесса [6].

Для данных из персональных аккаунтов социальных сетей характерны проблемы неправдивости, непригодности для анализа, устарелости или неполноты [7, 8], а также слабоструктурированный или неструктурированный характер данных с синтаксическими ошибками, аномалиями и смешанной природой, что затрудняет их анализ. При моделировании сообществ и информационного влияния адекватность построенных моделей зависит от качества и характера исходных данных, которые могут изме-

няться в процессе динамики интернет-сообщества [9]. Исходя из специфики первичных данных из соцсетей, возможным подходом к улучшению их качества является восстановление (импутация, missing values imputation) пропущенных значений и обогащение (data enrichment) недостаточно информативных слабоструктурированных показателей.

Согласно классификации, предложенной Литтлом и Рубином [10], множество методов обработки данных с пропусками можно разделить на четыре группы (рис. 1).

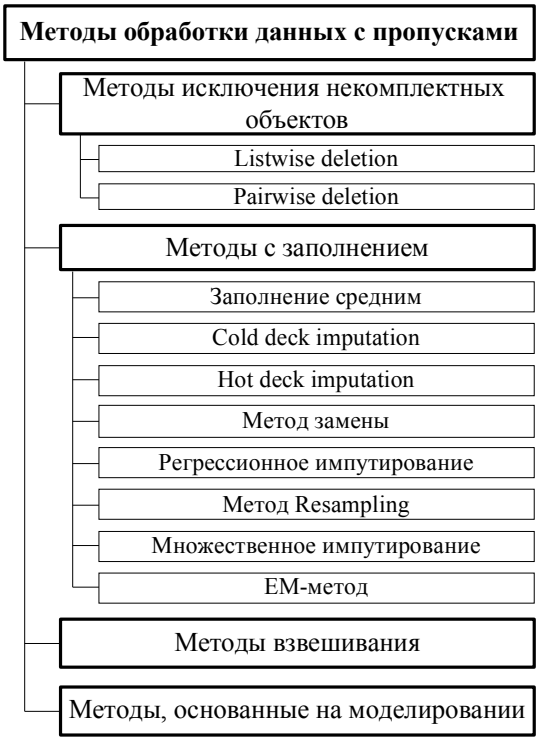

Рисунок 1 – Классификация методов обработки данных с пропусками

Первая группа содержит методы, реализующие процедуру исключения некомплектных объектов из анализа. *Вторая группа* включает методы, заполняющие пропуски, в результате применения которых получают «полные» данные, готовые к следующему этапу обработки. *Третья группа* состоит из методов рандомизированных выводов по данным выборочных обследований, построенных на весах плана, обратно пропорциональных вероятности выбора. *Четвертая группа* содержит методы, основывающиеся на построении модели порождения пропусков и функции правдоподобия.

Потенциально пропуски в данных могут отрицательно влиять на качество и надёжность результатов исследований. Между записями анализируемой таблицы могут существовать взаимосвязи, теряющиеся при удалении пропусков. При анализе данных из соцсетей, представленных в виде связного графа [11], удаление записей таблицы приводит к удалению его нод, вследствие чего теряется важная информация о характере взаимосвязей между вершинами в сети и их ролью. В таком случае оправдано применение методов обработки некомплектных данных, принадлежащих второй группе.

Для устранения пропусков в слабоструктурированных данных низкого качества тяжело подобрать единый метод обработки. В этом случае целесообразно применение методов, основанных на нескольких моделях [12–14]. Преимуществом таких подходов является то, что на каждом из этапов манипуляции с данными, возможно учесть их особенности и в случае получения неудовлетворительных результатов скорректировать алгоритм обработки.

Оценка пропущенных значений, как любой метод анализа, требует понимания их природы. Существует три вида механизмов формирования пропусков [10]: полностью случайные (missing completely at random, MCAR), случайные (missing at random, MAR) и неслучайные (missing not at random, MNAR). Для MCAR характерна независимость некомплектной переменной от собственных и любых других значений анализируемой таблицы. MAR предполагает зависимость некомплектной переменной от других переменных в таблице. MNAR имеет место, когда пропущенные переменные зависят от своих же значений [13]. Для данных из социальных сетей, между которыми априори существует явная зависимость (возраст и семейное положение, количество друзей и пол, семейное положение и пол) или может существовать неявная, характерны неслучайные пропуски, имеющие схему MAR или MNAR. В этом случае оправдано применение методов импутации некомплектных значений.

Ещё один фактор, влияющий на качество анализа данных – их информационная насыщенность. Применение методов, направленных на обогащение данных новой информацией, позволяет сделать их более ценными и значимыми с точки зрения решаемой аналитической задачи. Обогащение данных разделяют на два вида: внутреннее (повышение информативности данных за счет изменения их организации) и внешнее (привлечение дополнительной информации из внешних источников) [6]. Основными подходами к решению задачи внутреннего обогащения являются регрессионный метод и семантическое обогащение. Первый основан на построении линий регрессии (например, квантильной или байесовской) и позволяет расширить имеющийся набор данных [15]. Второй включает методы обогащения слабоструктурированной информации, наибольший интерес из которых для работы с данными из соцсетей представляют технологии Semantic Web [16]. Методы обогащения применяются, например, в случаях непригодности для анализа в первоначальном виде входящей в систему информации [17], необходимости одновременного обобщения малозначимых и выделения выдающихся качеств или структурных характеристик [18].

В контексте социальных сетей задача внешнего обогащения связана с восстановлением пропусков и решаема в силу наличия информационного следа, явных и неявных связей между исследуемыми объектами, доступностью методов API и др. факторов. Задача внутреннего обогащения актуальна при сла-

бой структурированности данных и решается за счет их реорганизации и включения в набор полезной информации, которая отсутствует в явном виде, но может быть получена из имеющихся данных [6].

Исходя из наличия инструментов для создания сообществ по интересам и коммуникационных функций социальных сетей, очевидно, что схожие пользователи посредством информационных связей объединяются в некоторые группы, что предполагает существование природных сегментов. Неоднородная связная структура пользователей социальных сетей и взаимосвязи между хранимыми данными дают возможность не только улучшить их качество, но и упростить этот процесс, применяя процедуру сегментации и выполняя обработку информации в отдельных сегментах со сниженной размерностью. Одним из возможных подходов к реализации такого процесса обработки, по мнению авторов, является комплексный взгляд на её решение путем синтеза ансамбля моделей на основе методов импутации и обогащения некачественных данных [19] с применением их предварительной сегментации. В работе предпринята попытка показать на эксперименте особенности структуры первичных данных из социальных сетей и возможность для улучшения их качества, с учетом их сетевой структуры.

Цель работы – обоснование подхода к улучшению качества исходных данных из социальной сети путём комплексного применения моделей сегментации, импутации и обогащения данных.

МАТЕРИАЛ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ. Результаты рекогносцировочных экспериментов в рамках проекта по моделированию сообществ потенциальных абитуриентов ВУЗа в г. Кременчуг и выявлению их лидеров [20] на базе социальной молодежной сети «ВКонтакте» показали низкое качество первичных данных. Основной причиной этого является отсутствие жестких правил относительно полноты заполнения информации персональных аккаунтов, что приводит к возникновению пропусков в генерируемых пользователями данных.

При решении любой аналитической задачи всё множество доступных для обработки информативных данных можно разделить на два вида: целевые данные и данные, не представляющие интерес для анализа. Например, при решении данной задачи [20], среди множества доступных из социальной сети показателей, можно выделить, указанные в табл. 1. При этом, целевые и нецелевые данные в свою очередь разделяются на 2 категории: характеристики пользователя и его личности, предоставляемые им самим (априорные) и характеристики его сетевой активности, вычисляемые и хранимые системой (апостериорные). Следует отметить тот факт, что все данные, хранимые системой, имеют показатель заполненности 100 %, т.е., по сути, не содержат пропусков.

Таблица 1 – Данные из социальной сети и процент их заполненности

Целевые факторы			Нецелевые факторы		
Название атрибута	Кол-во записей	Заполненность, %	Название атрибута	Кол-во записей	Заполненность, %
Дата рождения	33818	100	Уклон школы/класса	2303	6,81
Населённый пункт	33818	100	Факультет	4322	12,78
Семейное положение	13480	39,86	Кафедра	3192	9,44
Название среднего учебного заведения	9976	29,5	Партнер по отношениям	1840	5,44
Год окончания среднего учебного заведения	6706	19,83	Пол	33818	100
Название ВУЗа	5211	15,41			
Год окончания ВУЗа	3473	10,27			
Коммуникативность	33818	100	Номер домашнего телефона	8197	24,24
			Номер мобильного телефона	3876	11,46
			Учётная запись Skype	3142	9,29
			Учётная запись Facebook	68	0,2
			Учётная запись Twitter	91	0,27
			Учётная запись Livejournal	0	0
Комплектные данные					
Количество друзей	33818	100	Количество фотоальбомов	33818	100
Количество подписчиков	33818	100	Количество групп	33818	100
			Дата регистрации	33818	100
			Кол-во подписок	33818	100
Индекс посещаемости	33818	100	Время последнего посещения сети	33818	100
			Дата последнего посещения сети	33818	100
«Файловая активность»	33818	100	Кол-во видеозаписей	33818	100
			Кол-во аудиозаписей	33818	100
			Кол-во заметок	33818	100

			Кол-во фотографий	33818	100
			Кол-во фото «с пользователем»	33818	100
			Кол-во видео «с пользователем»	33818	100
Открытость	33818	100	Показатель возможности просмотра записей других пользователей	33818	100
			Показатель возможности написания сообщений пользователю	33818	100
			Показатель разрешения комментирования «стены»	33818	100
			Показатель возможности оставлять записи на «стене»	33818	100

Имея подобный набор данных, необходимо решить две важные задачи: отобрать для анализа наиболее значимые информативные данные и в случае их некомплектности применить методы заполнения пропущенных значений; попытаться улучшить качество малоинформативных данных, не имеющих пропусков, путём их обогащения.

Как видно из табл. 1, из целевых факторов для последующего анализа без процедуры предварительного заполнения пропусков доступны четыре информативных показателя: дата рождения, количество друзей, количество подписок и подписчиков. В процессе последующей обработки атрибут «дата рождения», не содержащий пропусков только в данном случае, поскольку является ключевым критерием поиска, был преобразован в показатель «возраст». Относительно данных о посещениях пользователем своей страницы, активности обмена файлами, настройках приватности и координатах вне социальной сети возможно применение методов их обогащения за счёт введения кодировок и подразделения на категории. Таким образом, данные о дате и времени последнего посещения сети пользователем, не несущие информационной нагрузки, преобразованы в показатель «индекс посещаемости»; данные о количестве аудио-, видеозаписей, фотографий и заметок – в показатель «файловая активность»; характеристики настроек приватности пользователя – в «открытость». Данные о наличии или отсутствии координат все социальной сети «Вконтакте» использованы для определения «коммуникативности» пользователя (табл. 1). Таким образом, в процессе синтеза модели сегментации на основе восьми описанных выше информативных показателей были проанализированы 33818 комплектных наборов данных о пользователях и выявлено наличие пяти сегментов (кластеров). Очевидно, что при таком подходе к отбору информативных факторов не учитывались данные, играющие немаловажную роль при решении задачи поиска сообществ потенциальных абитуриентов ВУЗа, но содержат пропуски, например, название среднего учебного заведения, его уклон, период обучения, год окончания, название высшего учебного заведения, год его окончания. В таком случае в целях повышения адекватности результатов моделирования актуальной становится задача восстановления значений целевых показателей.

Наличие сегментов, объединяющих схожих по характеристикам пользователей, и отличающихся от объектов из других кластеров, говорит о неоднородной структуре информации из социальной сети. Это значит, что в таблице «объект-признак» существуют связи между записями, описывающими объекты одного и того же сегмента. Присутствие таких связей делает возможным и обоснованным применение алгоритмов восстановления пропущенных значений.

С другой стороны, построение графа связей пользователей социальной сети предоставляет возможности для их описания в пространстве иных факторов: структуры взаимосвязей и информационного взаимодействия. Величина, описывающая неравномерность распределения связей между вершинами графа, называется его модулярностью (Modularity) [21]. Значения этого показателя, большие нуля, говорят о наличии подструктур (сообществ) с более тесными внутренними связями, чем во всём графе в целом. Другой показатель, характеризующий степень информационного взаимодействия некоторой вершины с остальными, и являющийся одним из ключевых при поиске лидеров [22] сообществ социальных сетей, называется мерой её центральности (Betweenness Centrality). Эта величина показывает, как часто через некоторую вершину проходят пути обмена данными в сети [23].

На рис. 2 в качестве примера структуры связей между пользователями социальной сети представлен граф G, построенный на основе данных, собранных для описанного выше эксперимента.

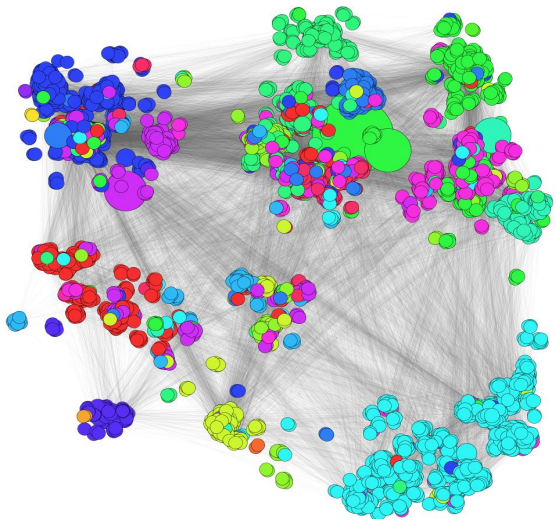

Рисунок 2 – Структура взаимосвязей между потенциальными абитуриентами

Визуализация графа выполнена с использованием программного обеспечения Gephi. Он состоит из 33818 вешин и уложен методом OpenOrd, позволяющим выявить группы нод с тесными информационными связями. Полученный граф характеризуется следующими показателями: диаметр $d(G)=10$, радиус $r(G)=1$, средняя степень вершин $deg(G)=26{,}4$, средняя длина кратчайшего пути $l(G)=3{,}2$. Исходя из структуры построенного графа и рассчитанного значения его модулярности $Q(G)=0{,}499$, можно предположить существование подграфов, объединяющих схожих по некоторым параметрам пользователей. Результаты применения метода поиска кластеров на графе, основанного на подсчёте его модулярности, подтвердили наличие 14 кластеров (вершины окрашены в соответствии с номером кластера, которому они принадлежат). Ранжирование вершин по значению Betweenness Centrality позволило выявить неоднородный характер их участия в информационных потоках сети. Выделение из общей массы нод, значение центральности которых явно преобладает перед остальной массой, говорит об их потенциальном лидерстве в сети. Отсюда можно предположить, что информативность каждой вершины графа также отличается, что необходимо учитывать при анализе и разработке методов восстановления или обогащения пропущенных значений в социальных сетях. Если учесть, что после совершения вершиной-лидером некоторого действия в пределах заданного отрезка времени существенное количество других пользователей повторяют это действие [24], то существует вероятность того, что это действие совершается вследствие большой схожести между характеристиками лидера и нод, наследующими его поведение.

В случае применения алгоритмов восстановления пропущенных значений в данных из социальной сети могут возникать вопросы относительно их устойчивости. Так как граф взаимосвязей между пользователями по своей природе – динамическая структура, то и в процессе перестройки моделей сообществ могут исчезать старые, появляться новые пользователи и связи между ними. Это проявляется в изменении таблицы «объект-признак», содержащей информацию о пользователях исследуемой сети – добавлении новых записей и удалении существующих.

Результаты экспериментов относительно устойчивости графа к внешним атакам сети показали, что этот параметр зависит от характера связей между вершинами. В случае воздействия на граф гомогенной структуры, ноды которого имеют приблизительно одинаковое количество связей, его скелетон начинает разрушаться при удалении случайным образом более 28% вершин. При воздействии на граф неоднородной природы с явно выделяющимися подструктурами, связность которых обеспечивают несколько нод с высоким показателем степени, и большинством вершин, имеющих незначительное количество связей, его структура начинает разрушаться при удалении более чем 75% нод [25]. Следовательно, для комплексных сетей такого типа оправдано применение алгоритмов восстановления пропущенных данных, поскольку при проведении атак на сеть и удалении из неё случайным образом вершин её общая структура сохраняется и остается устойчивой. Это значит, что сохраняются и внутренние связи между записями в таблице «объект-признак», что подтверждает возможность и целесообразность применения методов восстановления пропущенных значений.

В качестве эксперимента по восстановлению пропусков в данных из социальной сети авторами была предпринята попытка осуществить восстановление некомплектных целевых значений из описанного выше эксперимента. Для этого на начальном этапе была сформирована исходная таблица целевых данных «объект-признак» M, состоящая из 6 атрибутов (возраст, семейное положение, название ВУЗа, год окончания ВУЗа, название среднего учебного заведения, год окончания среднего учебного заведения) и 33818 записей. Для формирования комплектного набора данных матрица M с помощью применения метода Listwise deletion преобразована в модельную матрицу M', состоящую из 2682 записей. Далее в комплектной матрице в результате $n=6$ итераций поочередно сгенерированы 1, 5, 10, 20, 50 и 70 % пропусков, имеющих равномерное дискретное распределение, для каждого из атрибутов, и получены шесть матриц с пропусками M'_n. Для восстановления пропущенных значений за основу взят метод поиска ассоциативных правил с использованием алгоритма, предложенного в [14], поскольку они позволяют обнаружить закономерности и связи в наборах данных, характерных для социальных сетей.

Схема восстановления пропущенных значений имеет следующий вид. Каждая запись в матрице M'_n рассматривается как предметный набор, а номер записи о пользователе – как транзакция. В имеющемся наборе данных выполняется поиск ассоциативных правил, после чего пропущенные значения, обозначенные как «Missing», заполняются следующим образом. Множество правил разбивается на 6 групп в соответствии с атрибутами, содержащимися в следствиях, и сортируется в убывающем порядке по значению их достоверности. Для каждого атрибута вычисляется частота наиболее часто встречающегося значения, принимающаяся за пороговую величину при отборе используемых для восстановления ассоциативных правил. Все правила, достоверность которых ниже порогового значения, исключаются из рассмотрения. Далее для каждого из значений «Missing» выполняется поиск подходящего ассоциативного правила, содержащего в следствии не более чем одно значение, с наибольшей достоверностью.

Если такое правило находится, пропуск заполняется значением из следствия выбранного правила; если же правил, способных заполнить пропуск, не существует, значение «Missing» заполняется методом Most common value [26]. Результаты заполнения пропущенных значений для матриц M'_n представлены в табл. 2.

Таблица 2 – Процент верновосстановленных значений атрибутов

Процент пропусков	Восстанавливаемые атрибуты					
	Возраст (11)	Семейное положение (8)	Название школы (327)	Год окончания школы (40)	Название ВУЗа (217)	Год окончания ВУЗа (35)
1	31,38	27,27	9	36,36	45,45	36,36
5	28,18	28,18	10	27,27	42,73	26,36
10	25,57	28,31	7,31	21	37,44	25,11
20	25,51	25,97	6,83	25,06	51,25	25,97
50	14,59	25,89	8,38	12,67	49,23	16,32
70	12,50	27,21	8,46	12,04	48,83	15,43

Как видно из процентного соотношения верновосстановленных значений, такой алгоритм работает неудовлетворительно для данных из социальных сетей. В первую очередь, это обусловлено большим количеством уникальных значений (указано в скобках) таких атрибутов, как «Название школы» и «Название ВУЗа». При этом если для первого атрибута правила не генерируются из-за большого количества уникальных значений, и работает только метод Most common value, то для второго строятся правила, содержащие в следствии лишь наиболее часто встречающееся значение (с относительной частотой появления 49,36 %), при этом упускаются из вида редко встречающиеся.

Для снижения размерности значений рассматриваемых атрибутов и учёта скрытых связей между данными в качестве альтернативного подхода использовано восстановление пропущенных значений, которое выполнено с применением предварительной сегментации пользователей социальной сети. Так как процедура сегментации проводилась не в пространстве целевых факторов, требующих восстановления, а в пространстве атрибутов, не содержащих пропуски, из каждого найденного сегмента сформирована модельная матрица по схеме, описанной выше. Для дальнейшего исследования выбран сегмент № 1, содержащий наибольшее количество записей.

Таблица 3 – Количество записей в исходных и модельных сегментах

№ сегмента	0	1	2	3	4
Исходная матрица	1578	11550	8862	2120	9708
Модельная матрица	128	952	524	106	972

В результате применения выше описанного алгоритма восстановления пропущенных значений в сегменте № 1 получены следующие результаты (табл. 4):

Таблица 4 – Процент верновосстановленных значений атрибутов для сегмента № 1

Процент пропусков	Восстанавливаемые атрибуты					
	Возраст (6)	Семейное положение (8)	Название школы (222)	Год окончания школы (31)	Название ВУЗа (114)	Год окончания ВУЗа (22)
1	60	60	10	70	60	50
5	43,75	43,75	10,42	47,92	52,08	41,67
10	44,21	41,05	10,53	43,16	49,47	42,11
20	52,82	38,95	9,23	43,59	54,36	37,95
50	34,45	32,98	8,61	32,98	53,36	31,72
70	32,88	28,53	7,81	26,58	53	27,07

Результаты применения модифицированного метода восстановления пропущенных значений с использованием предварительной сегментации данных показали, что второй подход дает лучшие результаты для всех множеств модельных данных с различными процентами пропусков (рис. 3, 4). Как видно из эксперимента, применение ассоциативных правил внутри сегментов позволяет находить закономерности в данных и улучшать таким образом процент правильно восстановленных пропущенных значений даже с учётом того, что в данном случае рассматриваются результаты сегментации не в пространстве целевых для восстановления атрибутов.

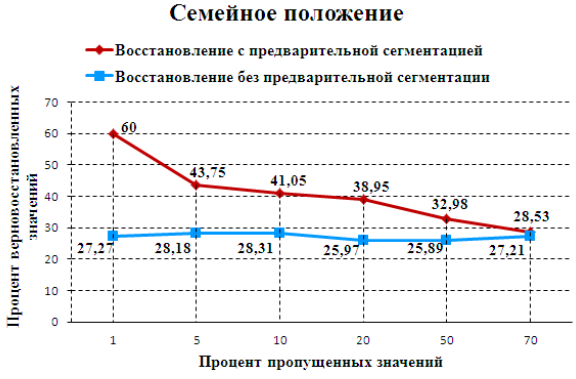

Рисунок 3 – Результаты восстановления пропущенных значений атрибута

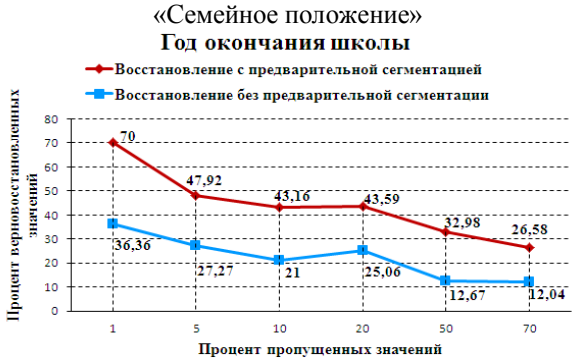

Рисунок 4 – Результаты восстановления пропущенных значений атрибута «Год окончания школы»

Таким образом, подтверждено наличие сегментов пользователей, имеющих схожие характеристики в пространстве априорно-апостериорных факторов, описывающих их и взаимосвязи между ними. Показано, что наличие таких гомогенных групп влияет на результаты импутации и позволяет повысить процент верно восстановленных значений. Полученные результаты дают основание для более глубокого исследования сегментации как одного из этапов предварительной обработки данных (в частности, в пространстве целевых для импутации информативных показателей) и разработки основ создания информационной технологии для автоматизированной процедуры улучшения качества исходных данных из социальных сетей.

ВЫВОДЫ. Сделан обзор и анализ существующих подходов к улучшению качества данных из соцсетей на основе заполнения пропущенных и обогащения существующих данных. На примере реальных данных из социальной сети выполнено исследование особенностей структуры информативных показателей и выявлены проблемы, влияющие на их пригодность к анализу.

Проведен эксперимент по восстановлению пропущенных значений с помощью применения комбинации методов ассоциативных правил и Most common value для всего множества данных о пользователях и отдельных сегментов. Дальнейшее развитие получил алгоритм импутации категориальных данных с использованием ассоциативных правил, который, в отличие от существующего, усовершенствован и адаптирован к применению на данных из социальной сети, атрибуты которых содержат множество уникальных значений, благодаря введению этапа предварительной сегментации. Показано, что неоднородность данных о пользователях социальных сетей и информационных связей между ними влияет на качество работы методов заполнения пропущенных значений. Применение предварительной сегментации позволило повысить процент верновосстановленных значений для атрибутов, имеющих порядка 30-ти уникальных значений с приблизительно одинаковой частотой встречаемости. Результаты восстановления атрибутов «возраст», «семейное положение», «год окончания школы», «год окончания ВУЗа» в среднем на 13,74–21,73 % больше, чем при использовании метода без предварительной сегментации.

Предложен подход на основе комбинации моделей первичной сегментации с последующим использованием ансамбля различных моделей импутации и обогащения данных, в зависимости от их характера и качества. Он может лечь в основу создания информационной технологии автоматизированной переработки информации с целью улучшения качества первичных данных из социальных сетей, что позволит повысить эффективность реализации DM-проектов в задачах моделирования сообществ соцсетей.

ЛИТЕРАТУРА

1. Freeman L.C. The Development of Social Network Analysis: A Study in the Sociology of Science. – Vancouver, CA: Empirical Press, 2004. – 205 p.

2. Dorogovtsev S., Mendes J. Evolution of networks // Advances in Physics. – 2002. – Iss. 51 (4). – PP. 1079–1187.

3. Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the Web // Computer Networks: The International Journal of Computer and Telecommunications Networking. – 2000. – Iss. 33. – PP. 309–320.

4. Watts D.J. The “New” Science of Networks // Annual Review of Sociology. – 2004. – Iss. 30. – PP. 243–270.

5. Bonchi F., Castillo C., Gionis A., Jaimes A. Social Network Analysis and Data Mining for Business Applications // Transactions on Intelligent Systems and Technology. – 2011. – Iss. 2. – P. 37.

6. Паклин Н.Б., Орешков И.И. Бизнес-аналитика: от данных к знаниям. – 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.

7. Bhagat S., Cormode G. Node classification in social networks // Social Network Data Analytics. – 2011. – PP. 115–148.

8. Lenhart A., Madden M. How teens manage their online identities and personal information in the age of My Space // Teens, Privacy and Online Social Networks, 2007. – Режим доступа: http://www.issuelab.org/resource/teens_privacy_and_online_social_networks_how_teens_manage_their_online_identities_and_personal_information_in_the_age_of_myspace. – Заголовок с экрана.

9. Rossi R., Neville J., Gallagher B., Henderson K. Role-dynamics: fast mining of large dynamic networks // Proceedings of the 21st international conference companion on World Wide Web, March 28–April 1, Hyderabad. – India, 2011. – PP. 997–1006.

10. Литтл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками / Пер. с англ. – М.: Финансы и статистика, 1990. – 336 с.

11. Coulon F. The use of Social Network Analysis in Innovation Research: A literature review // Proceedings of the DRUID Academy winter 2005 Ph.D. conference on Industrial and evolution and dynamics, January 27–29, Rebild. – Denmark, 2005. – PP. 1–28.

12. Reddy M.J.V., Kavitha B. Efficient ensemble algorithm for mixed numeric and categorical data // Proceedings of the IEEE International conference on Computational intelligence and computing research, December 28–29, Coimbatore. – India, 2010. – PP. 1–4.

13. Hlalele N., Nelwamondo F., Marwala T. Imputation of missing data using PCA, neuro-fuzzy and genetic algorithms // Proceedings of the 15th International Conference on Neural information processing of the Asia-pacific neural network assembly, November 25–28, Auckland. – New Zeland, 2008. – PP. 458–492.

14. Kaiser J. Algorithm for missing values imputation in categorical data with use of association rules // ACEEE International Journal on Recent Trends in Engineering & Technology. – 2011. – Iss. 6(1). – PP. 111–114.

15. Mount J. The Data Enrichment Method [Электронный ресурс]: – Режим доступа: http://www.win-vector.com/blog/2009/04/the-data-enrichment-method/. – Заголовок с экрана.

16. Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) // Искусственный интеллект и принятие решений. – Вып. 2008/1. – Москва, 2008. – С. 80–97.

17. Moraru A., Mladenic D. A framework for semantic enrichment of sensor data // Proceedings of the 34th International conference on Informational technology interfaces, June 25–28, Cavtat, Dubrovnik. – Croatia, 2012. – PP. 155–160.

18. Neun M., Weibel R., Burghardt D. Data Enrichment for Adaptive Generalisation // Proceedings of the 7th ICA Workshop on Generalisation and multiple representation, August 20–21, Leicester. – Great Britain, 2004. – PP. 1–6.

19. Слабченко О.О., Сидоренко В.Н. Ансамбли моделей восстановления и обогащения данных в задачах моделирования сообществ социальных сетей // Материалы II Международной научно-практической конференции «Полупроводниковые материалы, информационные технологии и фотовольтаика», 22–24 мая, г. Кременчуг. – Украина, 2013. – С. 224–226.

20. Слабченко О.О., Сидоренко В.Н., Пономарчук Р.А. Методы и алгоритмы выявления сообществ потенциальных абитуриентов и их лидеров в социальных сетях // Вестник Кременчугского национального университета. – 2013. – Вып. 1/2013 (78).– С. 53–61.

21. Clauset A., Newman M. E., Moore C. Finding community structure in very large networks // Physical Revue E. – 2004. – Iss. 70(6), 06611.

22. Hoppe B., Reinelt C. Social Network Analysis and the Evaluation of Leadership Networks // The Learedship Quarterly. – 2010. – Iss. 4. – PP. 600–619.

23. L. C. Freeman A set of measures of centrality based on Betweenness // Sociometry. – 1977. – Iss. 1. – PP. 35–41.

24. A. Goyal, F. Bonchi, Laks V. S. Lakshmanan Discovering leaders from community actions // Proceedings of the 17th ACM Conference on Information and Knowledge Management, October 26–30, Napa Valley, California. – USA, 2008. – PP. 499–508.

25. Albert R., Jeong H., Barabasi A.L. Error and attack tolerance of complex networks // Nature. – 2000. – Iss. 406. – PP. 378–382.

26. Grzymala-Busse J.W., Goodwin L.K., Grzymala-Busse W.J., Zheng X. Handling Missing Attribute Values in Preterm Birth Data Sets // Proceedings of the 10th international conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, August 31–September 2, Regina. – Canada, 2005. – PP. 342–351.

THE IMPROVEMENT OF INITIAL DATA QUALITY IN MODELING PLOBLEMS OF ONLINE COMMUNITIES ON THE BASE OF COMBINED IMPLEMENTATION OF SEGMENTATION, IMPUTATION AND DATA ENRICHMENT MODELS

O. Slabchenko, V. Sidorenko

Kremenchuk Mykhailo Ostrohradskyi National University

vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E–mail: vnsidorenko@gmail.com

A review and analysis of existing approaches to data quality improvement on the basis of imputation and data enrichment methods were done by the authors. Structural features of personal accounts's data from social networks were researched and problems that negatively affect their suitability for analysis were revealed. It is shown that presegmentation of the data reveals their inhomogeneous structure. A comparative missing values imputation experiment was done by a combination of association rules and most common value methods on the full data set and for a single segment. It was established experimentally that the use of imputation methods in respect to each segment separately allows increasing the percentage of correctly restored missing values. An approach to improvement data quality from social networks by a combined implementation of segmentation, imputation and data enrichment models was proposed.

Key words: complex networks, missing values imputation, data enrichment, Social Mining, Social Network Analysis.

REFERENCES

1. Freeman, L.C. (2004), *The Development of Social Network Analysis: A Study in the Sociology of Science*, Vancouver, CA: Empirical Press.

2. Dorogovtsev, S. and Mendes, J. (2002), "Evolution of networks", *Advances in Physics*, vol. 51, no. 4, pp. 1079–1187.

3. Broder, A., Kumar, R., Maghoul F., Raghavan, P., Rajagopalan, S., Stata, R, Tomkins, A., and Wiener, J. (2000), "Graph structure in the Web", *Computer Networks: The International Journal of Computer and Telecommunications Networking*, Iss. 33, pp. 309–320.

4. Watts, D.J. (2004), "The "New" Science of Networks", *Annual Review of Sociology*, Iss. 30, pp. 243–270.

5. Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. (2011), "Social Network Analysis and Data

Mining for Business Applications", *Transactions on Intelligent Systems and Technology*, iss. 2, p. 37.

6. Paklin, N.B. and Oreshkov, V.I. (2013), *Biznes-analitika: ot dannyh k znaniyam* [Business-analytics: From data to knowledge], Piter, SPb., Russia.

7. Bhagat, S. and Cormode, G. (2011), "Node classification in social networks", *Social Network Data Analytics*, pp. 115–148.

8. Lenhart, A. and Madden, M. (2007), "How teens manage their online identities and personal information in the age of My Space", *Teens, Privacy and Online Social Networks*, available at: www.issuelab.org/resource/teens_privacy_and_online_socil_networks_how_teens_manage_their_online_identities_and_personal_information_in_the_age_of_myspace (accessed September 18, 2013).

9. Rossi, R., Neville, J., Gallagher, B., and Henderson, K. (2011), "Role-dynamics: fast mining of large dynamic networks", *Proceedings of the 21st international conference companion on World Wide Web*, March 28–April 1, Hyderabad, India, pp. 997–1006.

10. Little, R.J. and Rubin, D.B. (1990) *Statisticheskiy analiz dannykh s propuskami* [Statistical analysis with missing data], Finansy i statistika , Moscow, Russia.

11. Coulon, F. (2005), "The use of Social Network Analysis in Innovation Research: A literature review", *Proceedings of the DRUID Academy winter 2005 Ph.D. conference on Industrial and evolution and dynamics*, January 27–29, Rebild, Denmark.

12. Reddy, M.J.V. and Kavitha, B. (2010), "Efficient ensemble algorithm for mixed numeric and categorical data", *Proceedings of the IEEE International conference on Computational intelligence and computing research*, December 28–29, Coimbatore, India, pp. 1–4.

13. Hlalele, N., Nelwamondo, F., and Marwala, T. (2008), "Imputation of missing data using PCA, neuro-fuzzy and genetic algorithms", *Proceedings of the 15th International Conference on Neural information processing of the Asia-pacific neural network assembly*, November 25–28, Auckland, New Zeland, pp. 458–492.

14. Kaiser, J. (2011), "Algorithm for missing values imputation in categorical data with use of association rules", *ACEEE International Journal on Recent Trends in Engineering & Technology*, vol. 6, no. 1, pp. 111–114.

15. Mount, J. (2009), "The Data Enrichment Method", available at: http://www.win-vector.com/blog/2009/04/the-data-enrichment-method/ (accessed October 02, 2013).

16. Khoroshevskiy, V.F. (2008), "Knowledge space in the Internet and Semantic Web (Part 1)", *Iskusstvennyy intellekt i prinyatiye resheniy*, vol. 2008, no. 1, pp. 80–97.

17. Moraru, A. and Mladenic, D. (2012), "A framework for semantic enrichment of sensor data", *Proceedings of the 34th International conference on Informational technology interfaces*, June 25–28, Cavtat, Croatia, pp. 155–160.

18. Neun, M., Weibel, R., and Burghardt, D. (2004)," Data Enrichment for Adaptive Generalisation", *Proceedings of the 7th ICA Workshop on Generalisation and multiple representation*, August 20–21, Leicester, Great Britain.

19. Slabchenko, O.O. and Sydorenko, V.N. (2013), "Ensembles of imputation and enrichment models in tasks of social networks' online communities modeling", *Materialy II Mezhdunarodnoy nauchno-prakticheskoy konferentsii «Poluprovodnikovyye materialy, informatsionnyye tekhnologii i foto-vol'taika»* [Proceedings of the 2nd International scientific-practical conference on Semiconductor materials, information technologies and photovoltaic], Kremenchug, May 22–24, 2013, pp. 224–226.

20. Slabchenko, O.O., Sydorenko, V.N. and Ponomarchuk, R.A. (2013), "Methods and algorithms for discovery the communities of potential entrances and their leaders in social networks", *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, vol. 1, no. 78, pp. 53–61.

21. Clauset, A., Newman, M. E., and Moore, C. (2004), "Finding community structure in very large networks", *Physical Revue E*, vol. 70, no. 6: 066111.

22. Hoppe, B., and Reinelt, C. (2010), "Social Network Analysis and the Evaluation of Leadership Networks", *The Leadership Quarterly*, iss. 4, pp. 600–619.

23. Freeman, L.C. (1977), "A set of measures of centrality based on Betweenness", *Sociometry*, iss. 1, pp. 35–41.

24. Goyal, A., Bonchi, F., and Lakshmanan, L.V.S. (2008), "Discovering leaders from community actions", *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, October 26–30, Napa Valley, USA, pp. 499–508.

25. Albert, R., Jeong, H., and Barabasi, A.L. (2000), "Error and attack tolerance of complex networks", *Nature*, iss. 406, p. 378–382.

26. Grzymala-Busse, J.W., Goodwin, L.K., Grzymala-Busse, W.J., and Zheng, X. (2005), "Handling Missing Attribute Values in Preterm Birth Data Sets", *Proceedings of the 10th international conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, August 31–September 2, Regina, Canada, pp. 342–351.

Стаття надійшла 09.10.2013.