УДК 004.04; 004.7

МЕТОДЫ И АЛГОРИТМЫ ВЫЯВЛЕНИЯ СООБЩЕСТВ ПОТЕНЦИАЛЬНЫХ АБИТУРИЕНТОВ И ИХ ЛИДЕРОВ В СОЦИАЛЬНЫХ СЕТЯХ

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

Предложена Data Mining-концепция выявления сообществ потенциальных абитуриентов в социальных сетях и их лидеров. Реализовано программное приложение, выполняющее сбор информации из социальных сетей в рамках построения экспериментальной модели поиска сообществ и их лидеров. Обоснована структура информативных показателей, описывающих потенциального абитуриента в сети, и осуществлена их редукция. Предложен подход к редукции пользователей в задаче отбора влиятельных агентов на основе последовательного применения моделей сегментации, что, в отличие от известных подходов, реализует более простую схему. Выполнен анализ и синтез алгоритма выделения лидеров сообществ путем редукции вершин их графа на основе ранжирования значений Betweenness Centrality. Установлено, что основная масса лидеров является членами сегмента высокоактивных многоконтактных пользователей с высоким рейтингом и входит в состав каждого из кластеров графа.

Ключевые слова: Social Network Analysis, Data Mining, социальные сети, поиск сообществ, агенты влияния.

МЕТОДИ ТА АЛГОРИТМИ ВИЯВЛЕННЯ СПІВТОВАРИСТВ ПОТЕНЦІЙНИХ АБУТУРІЄНТІВ І ЇХ ЛІДЕРІВ У СОЦІАЛЬНИХ МЕРЕЖАХ

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

Запропоновано Data Mining-концепцію виявлення співтовариств потенційних абітурієнтів у соціальних мережах і їх лідерів. Реалізовано програмний додаток, який виконує збір інформації із соціальних мереж у рамках побудови експериментальної моделі пошуку співтовариств і їх лідерів. Обґрунтовано структуру інформативних показників, які описують потенційного абітурієнта в мережі, і виконано їх редукцію. Запропоновано підхід до редукції користувачів у задачі відбору впливових агентів на основі послідовного застосування моделей сегментації, що, на відміну від відомих підходів, реалізує простішу схему. Виконано аналіз і синтез алгоритму виділення лідерів співтовариств шляхом редукції вершин їх графа на основі ранжування значень Betweenness Centrality. Установлено, що основна маса лідерів є членами сегменту високоактивних багатоконтактних користувачів з високим рейтингом і входить до складу кожного з кластерів графу.

Ключові слова: Social Network Analysis, Data Mining, соціальні мережі, пошук співтовариств, агенти впливу.

АКТУАЛЬНОСТЬ РАБОТЫ. Демографический кризис 1990–2000 гг. на постсоветском пространстве с одновременным ростом количества новых ВУЗов привели на сегодняшний день к крайнему снижению количества абитуриентов. В условиях жесткой конкуренции ВУЗы вынуждены разрабатывать и применять новые эффективные стратегии привлечения абитуриентов еще до старта приемной кампании. В современных условиях, при высоком уровне развития электронных средств коммуникации и социальных сетей, инновационным подходом к взаимодействию с абитуриентами является использование информационных технологий для проведения результативной агитационной кампании.

Анализ существующих подходов к решению вышеупомянутой проблемы показал наличие слабого звена в этом вопросе – выявление целевой аудитории и снижение затрат на проведение агитационных мероприятий [1]. Возникает задача создания такой системы, которая могла бы автоматически осуществлять сбор данных о потенциальных абитуриентах, выделять целевую аудиторию и выявлять её лидеров с целью разработки оптимальной стратегии информационного взаимодействия с ними. Исходя из анализа современной литературы в области информационных технологий и междисциплинарных знаний, решение такой задачи возможно путём использования методов Data Mining (DM) [2] и Social Network Analysis (SNA) [3], позволяющих выявить и визуализировать явные и скрытые сообщества пользователей социальных сетей по интересующим нас признакам. Взгляд на данную задачу с точки зрения DM в конечном итоге дает возможность тиражирования знаний, обеспечивающего возможность пользователям, не разбирающимся в анализе данных, применять его результаты на практике.

Стремительное развитие социальных сетей и возможности сбора информации с них [4] средствами API [5] привело к ощутимому повышению интереса к SNA и появлению его новых методов [6], приобретающих всё большую популярность и применяющихся в различных сферах [7]: системы поиска экспертов [8] и сбор команд специалистов [9], социальные рекомендации [10], поисковые системы людей и документов [11], маркетинг [12], телекоммуникации [13], реклама [14] и многие другие. При этом SNA исследует структуру взаимоотношений между участниками вышеперечисленных прикладных областей путем обнаружения неявных связей между ними [15] с привлечением теории графов.

В рамках решения подобных задач возникают специфические проблемы, такие как поиск явных (*explicit*) и неявных (*implicit*) сообществ и лидеров. Сообщество можно определить как множество сущностей, в пределах которого связи между объектами сильнее, чем за его пределами. Выявление сообществ является важной проблемой, включающей в

себя классификацию нод-участников сети, и, как следствие, выявление однородных групп, групп лидеров или критических групп соединительных звеньев [16]. Сообщества могут соответствовать группам страниц в Интернете, имеющим похожие темы [17], группам связанных индивидуумов в социальных сетях [18] и т.д. Выявление сообществ фактически является аналогом кластеризации, традиционной задачи Data Mining, применительно к различным социальным сетям. Подходы к выделению целевых групп абитуриентов на основе выявления сообществ дают возможность их моделирования как единого целого, с последующим применением моделей информационного влияния и управления [19]. При этом поиск лидеров в сообществах является актуальной задачей, поскольку при исследовании или разработке моделей информационного влияния важно иметь данные о характере взаимодействия членов сообществ, связи между ними и закономерностях распределения информационных потоков. Согласно [20], некоторая нода является лидером, если после совершения ею определенного действия в пределах заданного отрезка времени существенное количество других пользователей повторяют это же действие. Проблемы выявления лидеров широко распространены во многих сферах, например, в [21] рассматривается «гипотеза влиятельных пользователей» применительно к задачам маркетинга; выбор множества индивидуумов для предложения какого-либо продукта или инновации [22]; распространение и максимизация влияния в конкурентной социальной сети и привлечение последователей, вирусный маркетинг [23]; распространение социального влияния [24] и т.д. В целом, поиск лидеров сообществ и использование различных моделей влияния применяется в случаях поиска целевой аудитории для предложения и продвижения какого-либо продукта. Таким образом, выявление сообществ и их лидеров в контексте проблемы взаимодействия с абитуриентами является актуальной задачей.

Цель работы – повышение эффективности информационного взаимодействия с потенциальными абитуриентами путём построения системы выявления и моделирования их сообществ и лидеров в социальных сетях.

МАТЕРИАЛ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ. Исходя из вышесказанного, авторы предлагают рассматривать совокупность абитуриентов в рамках иной парадигмы, отличной от традиционной как сложную взаимосвязанную неоднородную структуру, развивающуюся как единое целое, со своими явными и неявными связями, лидерами и динамикой. Такой взгляд выдвигает необходимость разработки концепции информационной системы, позволяющей выполнять поиск, сбор и обогащение информации из социальной сети о потенциальных абитуриентах, визуализировать связи между ними и выявлять их сообщества и лидеров. Предлагаемая система является DM-решением, включающим этапы сбора, предобработки, анализа и интерпретации полученных результатов с необходимостью дальнейшей разработки определенной стратегии информационного влияния (рис. 1).

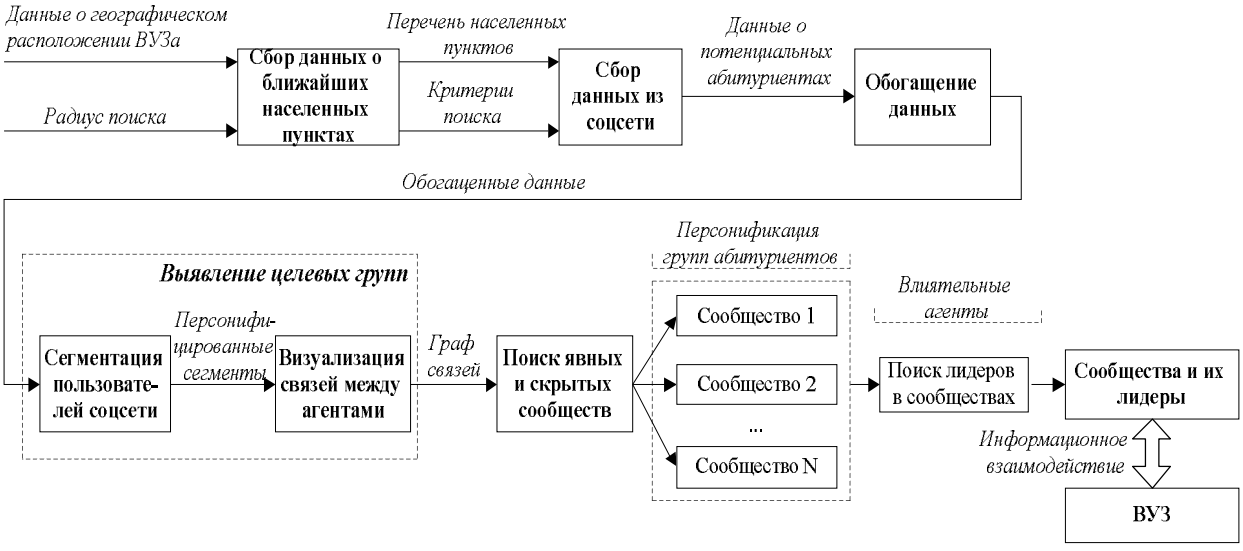

Рисунок 1 – Концепция системы выявления потенциальных абитуриентов и взаимодействия с ними в социальной сети

Отправная точка для сбора информации и построения модели – показатели географического положения ВУЗа и задание радиуса поиска. Данные, собираемые из соцсети, имеют характерные особенности – высокую размерность, большой объём и наличие пропусков. Поэтому возникает задача их обоснования, отбора и редукции перед началом построения моделей. В нашем случае последняя проблема решается путём применения последовательной редукции исходных данных – выделение целевых групп пользователей соцсети с последующей кластеризацией графа связей между отобранными пользователями (поиск явных и неявных сообществ). В результате выделяются группы наиболее вероятных абитуриентов, внутри которых происходит поиск лидеров (наиболее влиятельных

агентов). Такой подход, используемый в директ-маркетинге, решает важную задачу: минимизацию затрат на взаимодействия с малоактивными абитуриентами, не откликающимися на агитацию [25]. Таким образом, отсеивание пользователей, от которых нет отклика, приводит к снижению затрат на различные агитационные мероприятия в десятки раз, при этом обеспечивая взаимодействие с незначительно изменившимся количеством потенциально заинтересованных абитуриентов.

Построение экспериментальной модели поиска сообществ и их лидеров осуществлено на основе данных о потенциальных абитуриентах из социальной сети «Вконтакте», являющейся по данным [26] самой популярной среди украинской молодежи. Механизм сбора реализован посредством разработки приложения, взаимодействующего с сервером «Вконтакте» через предоставляемые методы API. Основные критерии поиска – возраст пользователя (от 16 до 25 лет) и место его проживания (г. Кременчуг). В основе работы системы, собирающей данные, лежит такая идея: информационные показатели, описывающие пользователей соцсетей, делятся на два типа – априорные и апостериорные [27]. Первые содержат информацию о пользователе, его образовании, семейном положении, хобби и т.д., вторые – определяют его активность в сети и интенсивность взаимодействия с другими пользователями. Очевидно, для полноты и адекватности анализа характера взаимосвязей между пользователями необходимо учитывать оба типа данных. Исследование их структуры показало, что отличительной особенностью информации из соцсети является наличие пропусков, поскольку не существует жестких правил относительно полноты заполнения информации на странице пользователя. Поэтому на первом этапе отбора данных были определены их типы и проведена оценка полноты по десятибалльной шкале. Для анализа были отобраны показатели, позволяющие отнести объект к целевой группе или выявить его высокую активность в сети, не имеющие пропущенных значений. Снижение размерности выполнено с помощью объединения групп сходных информативных данных в общий показатель путём суммирования их значений. Структура отобранных для дальнейшего анализа показателей, представленная на рис. 2, является основой для применения в дальнейшем методов интеллектуального анализа данных для решения задачи сегментации потенциальных абитуриентов с целью определения целевых групп и поиска их лидеров.

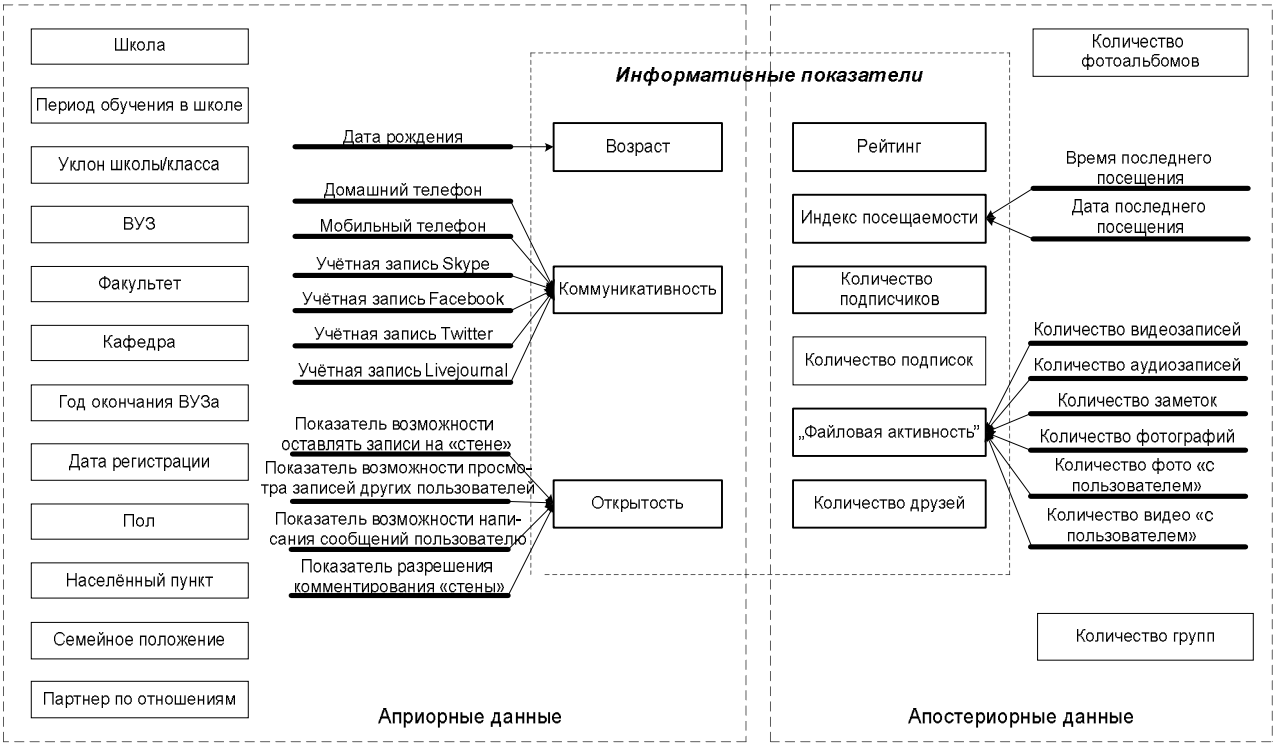

Рисунок 2 – Структура априорно-апостериорных данных о пользователях и информативных показателях

С учетом эффективности вычислительной процедуры для больших выборок, высокой размерности данных и количественной природы пространства признаков, сегментация потенциальных абитуриентов проведена с помощью карт Кохонена, имеющих существенное преимущество перед рядом других методов – возможность проецирования рельефа многомерных данных на двумерную плоскость. Синтез модели сегментации выполнен путём применения метода *k*-средних с евклидовой метрикой в пространстве описанных выше показателей (рис. 2) на базе аналитической платформы Deductor Studio Academic 5.2 [28]. Анализ результатов сегментации показал наличие пяти сегментов потенциальных абитуриентов (рис. 3,а). Основными показателями, по которым прослеживается различие между членами сегментов, являются возраст (рис. 3,б), рейтинг (рис. 3,в) и индекс посещаемости соцсети (рис. 3,г).

Выявленные сегменты были персонифицированы следующим образом: 0-ой сегмент (6,3 %) – «Высокоактивные многоконтактные пользователи с высокий рейтингом», 1-й сегмент (25,9 %) – «Высокоактивная среднеконтактная молодежь», 2-й сегмент (34,6 %) – «Старшая возрастная группа с большим количеством друзей», 3-й сегмент (4,7 %) – «Популярные высокоактивные пользователи», 4-й сегмент (28,5 %) – «Низкоактивные пользователи с невысоким рейтингом».

Для отбора потенциальных абитуриентов, подлежащих первоочерёдному анализу, были выделены пользователи возрастом 16–18 лет. Исходя из результатов персонификации сегментов, можно увидеть, что основная масса абитуриентов данного возраста входит в состав первого сегмента. Все остальные сосредоточены в первом, третьем и четвертом, а во второй сегмент данная возрастная категория не попала совсем (рис. 4).

Поэтому при выделении целевой аудитории можно сделать следующие выводы:

1. Для анализа целесообразен отбор всех представителей первого сегмента возрастом 16–18 лет.
2. Особое внимание нужно обратить на пользователей данной возрастной группы, попавших в состав нулевого и третьего сегментов, поскольку они наиболее активны, имеют большое количество подписчиков и мультимедийных данных, и, соответственно, интенсивно взаимодействуют с другими пользователями соцсети.
3. Пользователей в возрасте 16–18 лет, попавших в четвертый сегмент, не стоит принимать во внимание по причине их низкой активности.

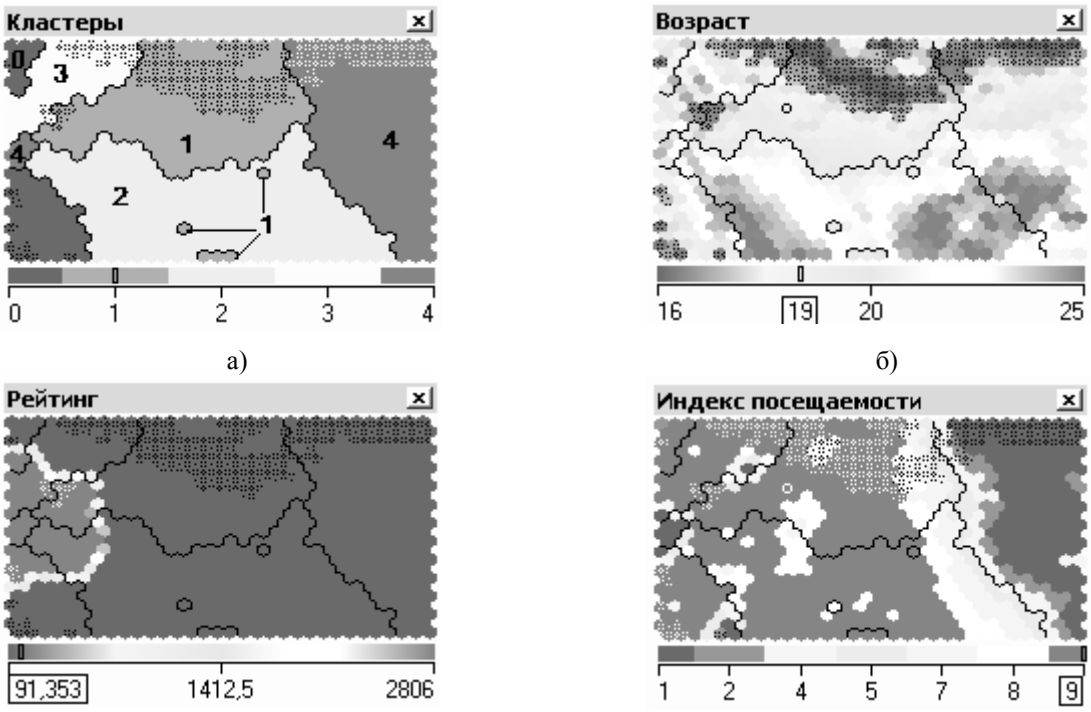

Рисунок 3 – Результаты сегментации и проецирование на плоскости основных дифференцирующих факторов: – группа потенциальных абитуриентов в возрасте 16–18 лет

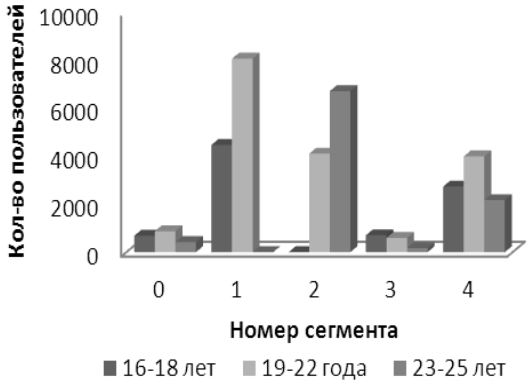

Рисунок 4 – Распределение возрастных групп пользователей по сегментам

Для выявления наиболее активных участников сетевого сообщества с точки зрения их информационного взаимодействия с другими пользователями для определённых выше целевых групп целесообразно построить граф взаимосвязей.

Процедура синтеза модели выявления сообществ абитуриентов была реализована в среде программного обеспечения Gephi для членов нулевого, первого и третьего сегментов для возраста 16–18 лет (рис. 5). В результате был получен граф из 5542 вершин, имеющих ненулевую степень. Для его укладки использован метод OpenOrd [29], применяющий алгоритмы Force-directed [30] и Average-link clustering [31] и визуализирующий большой граф таким образом, что соседние вершины притягиваются, а несвязанные друг с другом, наоборот, отталкиваются. Как видно из рис. 5, на полученном графе

выделяются ячейки с более тесными внутренними связями, чем во всем графе в целом. Поэтому можно предположить существование подграфов, представляющих собой группы пользователей с похожими параметрами.

Для их поиска был применен метод разбиения графа, основанный на расчёте значения модулярности (Q) [16], которое показывает, насколько распределение дуг данного графа отклоняется от равномерного. Пусть c_i обозначает членство вершины v_i в сообществе, тогда

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j), \qquad (1)$$

где $\frac{k_i k_j}{2m}$ – ожидаемое число дуг между нодами i и j в модели графа с равномерным распределением, m – количество дуг в графе, A_{ij} – реальное число дуг между i и j, $\delta(c_i, c_j)=1$ если $c_i = c_j$.

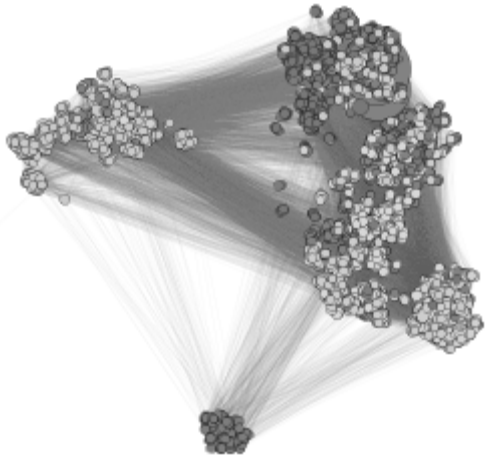

Рисунок 5 – Граф взаимосвязей между кластерами потенциальных абитуриентов

Полученные результаты доказали наличие восьми кластеров с модулярностью графа, равной 0,426.

При построении графа ранжирование вершин выполнялось по значению Betweenness Centrality (Bc) – величины, которая характеризует центральность вершины графа, и равна количеству кратчайших путей от всех вершин в остальные другие, проходящих через данную:

$$Bc(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \qquad (2)$$

где σ_{st} – общее количество кратчайших путей от ноды s к ноде t, $\sigma_{st}(v)$ – количество таких путей, которые проходят через ноду v [32].

Взяв за основу модель поиска лидеров сети на базе расчета Betweenness Centrality, был выполнен синтез алгоритма выделения лидеров групп сообществ на графе (рис. 6).

Эта характеристика является одной из ключевых при поиске лидеров [33], т.к. чем чаще через некоторую вершину проходят пути обмена данными в сети, тем выше степень её информационного взаимодействия с другими нодами. Путем расчета Betweenness Centrality также обычно проводится нахождения мостов (ключевых вершин, соединяющих между собой вершины) сети [34]. Следовательно, чем больше значение Bc имеет вершина, тем активнее она принимает участие в информационных потоках сети.

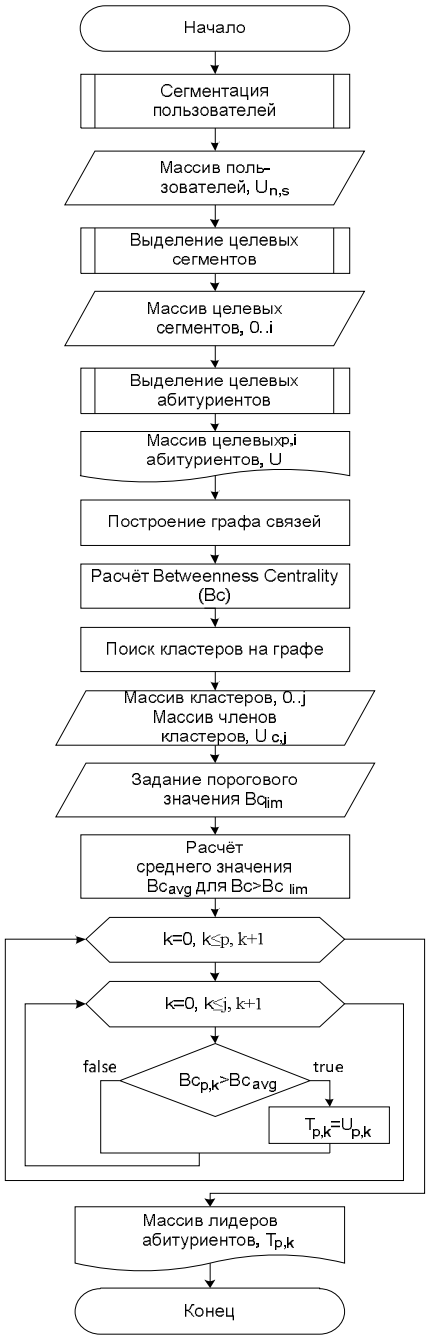

Рисунок 6 – Блок-схема алгоритма нахождения лидеров на основе расчёта Betweenness Centrality

Данный алгоритм основан на последовательном применении ансамбля моделей редукции данных:

сегментации всех пользователей сети, данные о которых были собраны на предыдущем этапе; персонификации полученных сегментов и выделении целевых групп; построении графа связей, его кластеризации и выявлении сообществ абитуриентов, и, наконец, выявлении лидеров сообществ абитуриентов путём расчёта значения Betweenness Centrality для каждой вершины. Данный алгоритм при поиске лидеров сообществ требует задания порогового значения Bc_{lim}, при котором вершины, имеющие такую и большую величину, идентифицируются как лидеры. В данном случае пороговое значение Betweenness Centrality было задано равным среднему значению Bc_{avg} по всему графу.

В результате применения предложенного алгоритма был получен редуцированный граф лидеров с 223-мя вершинами (рис. 7). На нём представлены члены трех сегментов, найденных в результате построения модели сегментации потенциальных абитуриентов на основе самоорганизующихся карт Кохонена.

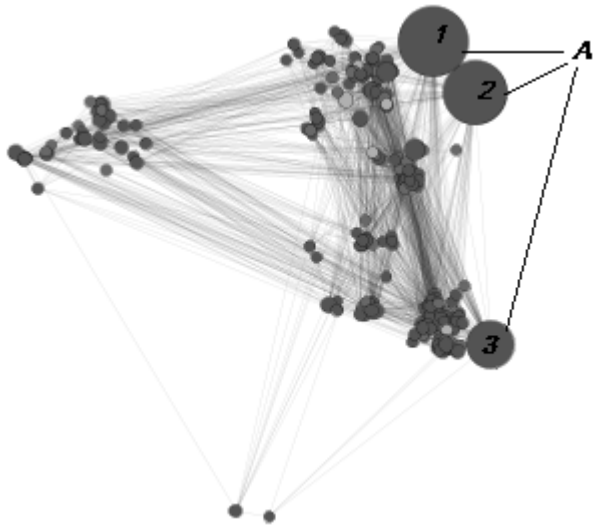

Рисунок 7 – Редуцированный граф связей между 223 лидерами сети из 5542 пользователей

Основную массу лидеров составляют члены нулевого сегмента (74,44 %) – высокоактивные многоконтактные пользователи с высоким рейтингом, а значит, именно в нем целесообразно искать агентов влияния для информационного воздействия. Поскольку представители этого сегмента входят в состав каждого из кластеров на графе, посредством информационного взаимодействия с ними можно охватить основную часть вершин графа всех потенциальных абитуриентов. Члены первого сегмента представляют 21,97 % лидеров, третьего – 3,59 %.

Анализ отдельных вершин показал, что три из них с наибольшим значением Betweenness Centrality представляют две учётные записи (точки 1 и 2) одного и того человека (точка A), проявляющего очень высокую активность и степень взаимодействия со своими многочисленными друзьями в соцсети на обеих своих страницах, а третья – группу по интересам, которую он администрирует. При этом первые две ноды и третья при кластеризации графа отнесены в разные сегменты, что говорит об их представлении различных интересов и разном круге взаимодействия с остальными пользователями.

Таким образом, построенная модель выделения лидеров сообществ позволяет не только определить наиболее влиятельных и активных потенциальных абитуриентов в соцсетях, но и выполнять визуализацию и анализ связей между ними. Полученные результаты дают основание для реализации и исследования модели, в основе которой будут лежать результаты персонификации каждого из найденных кластеров на графе, а также детальный анализ выявленных лидеров. Подобные нетривиальные знания о составе участников выявленных сообществ по интересам и другим объединяющим характеристикам помогут построить их профили с целью разработки оптимальной и эффективной стратегии взаимодействия с каждой из групп абитуриентов с применением индивидуального подхода.

ВЫВОДЫ. Обоснована Data Mining-концепция системы обнаружения сообществ потенциальных абитуриентов и их лидеров на основе информационного подхода, которая, в отличие от известных, позволит автоматически проводить сбор и анализ информации об абитуриентах из соцсетей, строить модели их скрытых сообществ, выявлять агентов влияния с целью повышения эффективности проведение агитационных мероприятий.

Реализовано программное приложение, выполняющее сбор информации из соцсети в рамках построения экспериментальной модели поиска сообществ и их лидеров. Обоснованы количество и структура информативных показателей, и выполнено снижение их размерности для решения задачи сегментации абитуриентов.

На основе последовательного применения моделей сегментации предложен подход, состоящий из этапов выделения целевой аудитории с помощью карт Кохонена и поиска сообществ абитуриентов путём кластеризации графа связей между ними, что, в отличие от известных подходов, позволяет реализовать упрощенную схему редукции исходных данных в задаче отбора агентов влияния в модели социальных сообществ потенциальных абитуриентов.

Выполнен анализ и синтез алгоритма нахождения лидеров групп сообществ на основе редукции вершин графа, ранжированных по значениям Betweenness Centrality, в основе которого лежит исключение вершин, имеющих значение показателя ниже заданного порога.

Экспериментальное исследование показало, что 74,44 % найденных лидеров сообществ являются членами нулевого сегмента высокоактивных многоконтактных пользователей с высоким значением показателя «Рейтинг» и представлены во всех кластерах графа взаимосвязей между абитуриентами.

ЛИТЕРАТУРА

1. Вертоградов В.А. CRM для ВУЗов: проведение маркетинговых мероприятий для абитуриентов [Электронный ресурс]// Информационно-коммуникационные технологии в образовании. – 2011. – Режим доступа: http://www.ict.edu.ru/news/study/4235/ – Заголовок с экрана

2. Han J., Kamber M. Data Mining: Concepts and Techniques. – Morgan Kaufmann, 2011. – 744 p.

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

4. Winkler W.E. Methods for evaluating and creating data quality // Information Systems. – 2004. – Iss. 7. – PP. 531–550.

5. Documentation VK API [Электронный ресурс]: – Режим доступа: http://vk.com/developers.php#devstep2. – Заголовок с экрана.

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

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

8. Yimam-Seid D., Kobsa A. Expert-Finding Systems for Organizations // Problem and Domain Analysis and the DEMOIR Approach. – 2003. – Iss. 1. – PP. 1–24.

9. Lappas T., K. Liu E. Terzi Finding a team of experts in social networks // Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, June 28–July 1, Paris. – France, 2009. – PP. 467–476.

10. Schifanella R., Barrat A., Cattuto C., Markines B., Menczer F. Folks in Folksonomies: social link prediction from shared metadata // Proceedings of the third ACM international conference on Web search and data mining, February 4–6, 2010, New York. – USA, 2010. – PP. 271–280.

11. Ronen I., Shahar E., Sigalit U. et al. Social networks and discovery in the enterprise // Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, July 19–23, 2009, Boston. – USA, 2009. – P. 836.

12. Bharathi S., Kempe D., Mahyar S. Competitive influence maximization in social networks // Proceedings of the 3rd international conference on Internet and network economics, December 12–14, 2007. – San Diego, USA, 2007. – PP. 306–311.

13. Phithakkitnukoon S., Dantu R. Adequacy of data for characterizing caller behavior // Proceedings of the 13th ACM SIGKDD Intarnational Conference of Knowledge Discovery and Data Mining, August 24. – Las Vegas, USA, 2008.

14. Provost F., Dalessandro B., Hook R., Zhang X., Murray A. Audience selection for on-line brand advertising: privacy-friendly social network targeting // Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, June 28–July 1, Paris. – France, 2009. – PP. 707–716.

15. Ehrlich K., Carboni I. Inside Social Network Analysis // IBM Watson Research Center. – New York, USA, 2005. – Technical Report 5–10.

16. Coscia M., Giannotti F., Pedreschi D. A classification for community discovery methods in complex networks // Statistical Analysis and Data Mining. – 2011. – PP. 512–546.

17. Flake G.W., Lawrence S., Giles C.L., Coetzee F.M. Self-organization and identification of Web communities // Computer. – 2002. – Iss. 3. – PP. 66–70.

18. Girvan M., Newman M.E. Community structure in social and biological networks // Proceedings of the National Academy of Sciences of the United States of America. – 2002. – Iss. 12. – PP. 7821–7826.

19. Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Социальные сети: модели информационного влияния, управления и противоборства: монография. – Москва, 2010. – 225 с.

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

21. Watts D.J., Dodds P.S. Influentials, Networks, and Public Opinion Formation // Journal of consumer research. – 2007. – Iss. 4. – PP. 441–458.

22. Kempe D., Kleinberg J., Tardos É. Maximizing the spread of influence through a social network // Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington. – USA, 2003. – PP. 137–146.

23. Carnes T., Nagarajan R., Wild S.M., Van Zuylen A. Maximizing influence in a competitive social network: a follower's perspective // Proceedings of the ninth international conference on Electronic commerce, Minneapolis. – USA, 2007. – PP. 351–360.

24. Dodds P.S., Watts D.J. A generalized model of social and biological contagion // Journal of theoretical biology. – 2005. – Iss. 4. – PP. 587–604.

25. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009. – 624 с.

26. В социальных сетях зарегистрировано 30 млн украинских аккаунтов [Электронный ресурс]. – Режим доступа: http://ain.ua/2012/09/21/96039 – Заголовок с экрана.

27. Сидоренко В.М., Слабченко О.О., Морванюк В.О. Сегментація користувачів Інтернет кафедри університету на основі апріорно-апостеріорної інформації в задачах моніторингу інтернет-активності // Вісник Кременчуцького національного університету. – Вип. 5/2011 (70). – Кременчук, 2011. – С. 52–54.

28. Описание платформы Deductor [Электронный ресурс]. – Режим доступа: http://www.basegroup.ru/deductor/description – Заголовок с экрана.

29. Martin S., Brown W.M., Klavans R., Boyack K.W. OpenOrd: an open-source toolbox for large graph layout // Proceedings of the SPIE 7868, Visualization

and Data Analysis. – San Francisco, California: USA, 2011.

30. Force Directed Layouts [Electronic resource]. – Mode of access: http://philogb.github.com/blog/2009/09/30/force-directed-layouts – Scr. title.

31. Single-Link, Complete-Link & Average-Link Clustering Layouts [Electronic resource]. – Mode of access: http://nlp.stanford.edu/IR-book/completelink.html – Scr. title.

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

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

34. Freeman L.C. Centrality in social networks: conceptual clarification // Social Networks. – 1979. – Iss. 1. – PP. 215–239.

METHODS AND ALGORITHMS FOR DISCOVERY THE COMMUNITIES OF POTENTIAL ENTRANCES AND THEIR LEADERS IN SOCIAL NETWORKS

O. Slabchenko, V. Sidorenko, R. Ponomarchuk

Kremenchuk Mykhailo Ostrohradskyi National University

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

Data Mining-conception of potential entrances' communities and their leaders' discovery in social networks is offered. The software tool for data collection from a social network within an experimental model of communities searching and their leaders is implemented. The structure of informative characteristics that describes user in a network is based. The reduction of potential entrances is performed. The approach to users' reduction in a task of influential actors' selection on a base of cascade application of segmentation models is offered. As opposite to well-known methods, this approach realizes an easier scheme. The analysis and synthesis of community leaders separation is made by the reduction of vertices of their graph on the base of Betweenness Centrality values ranking. It is determined that the main part of the leaders is members of multi-contact segment of users with high rating and enters into the composition of each cluster in a graph.

Ключевые слова: Social Network Analysis, Data Mining, social networks, communities' discovery, influence agents.

REFERENCES

1. Vertogradov V.A. CRM for Universities: implementation of marketing activities for entrants // Informacionno-kommunikacionnye technologii v obrazovanii. – 2011. – Mode of access: http://www.ict.edu.ru/news/study/4235/ – Scr. title. [in Russian]

2. Han J., Kamber M. Data Mining: *Concepts and Techniques,* 3rd Edition: Morgan Kaufmann, 2011. – 744 p.

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

4. Winkler W.E. Methods for evaluating and creating data quality // *Information Systems*. – 2004. – Iss. 7. – PP. 531–550.

5. *Documentation VK API* [Online resource]: – Mode of access: http://vk.com/developers.php#devstep2 – Scr. title. [in Russian]

6. Watts D.J. The "New" Science of Networks // *Annual Review of Sociology*. – 2004. – Iss. 30. – PP. 243–270.

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

8. Yimam-Seid D., Kobsa A. *Expert-Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach.* – 2003. – Iss. 1. – PP. 1–24.

9. Lappas T., Liu K., Terzi E. Finding a team of experts in social networks // *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, June 28–July 1, Paris, France, 2009. – PP. 467–476.

10. Schifanella R., Barrat A., Cattuto C., Markines B., Menczer F. Folks in Folksonomies: social link prediction from shared metadata // *Proceedings of the third ACM international conference on Web search and data mining,* February 4–6, 2010, New York, USA. – PP. 271–280.

11. Ronen I., Shahar E., Sigalit U. et al. Social networks and discovery in the enterprise // *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, July 19–23, 2009, Boston, USA. – P. 836.

12. Bharathi S., Kempe D., Mahyar S. Competitive influence maximization in social networks // *Proceedings of the 3rd international conference on Internet and network economics*, San Diego, USA, December 12–14, 2007. – PP. 306–311.

13. Phithakkitnukoon S., Dantu R. Adequacy of data for characterizing caller behavior // *Proceedings of the 13th ACM SIGKDD Intarnational Conference of Knowledge Discovery and Data Mining*, Las Vegas, USA, August 24, 2008.

14. Provost F., Dalessandro B., Hook R., Zhang X., Murray A. Audience selection for on-line brand advertising: privacy-friendly social network targeting // *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, June 28–July 1, Paris, France, 2009. – PP. 707–716.

15. Ehrlich K., Carboni I. *Inside Social Network Analysis* // IBM Watson Research Center. – New York, USA, 2005. – Technical Report 5–10.

16. Coscia M., Giannotti F., Pedreschi D. A classification for community discovery methods in complex networks // *Statistical Analysis and Data Mining*. – 2011. – PP. 512–546.

17. Flake G.W., Lawrence S., Giles C.L., Coetzee F.M. Self-organization and identification of Web communities // *Computer*. – 2002. – Iss. 3. – PP. 66–70.

18. Girvan M., Newman M.E. Community structure in social and biological networks // *Proceedings of the National Academy of Sciences of the United States of America*. – 2002. – Iss. 12. – PP. 7821–7826.

19. Gubanov D.A., Novikov D.A., Chartishvili A.G. *Socialnye seti: modeli informacionnogo vliyaniya, upravleniya i protivobotstva* [Social networks: informational influence, management and contention models]: monograph. – Moscow, 2010. – 225 p. [In Russian]

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

21. Watts D.J., Dodds P.S. Influentials, Networks, and Public Opinion Formation // *Journal of consumer research*. – 2007. – Iss. 4. – PP. 441–458.

22. Kempe D., Kleinberg J., Tardos É. Maximizing the spread of influence through a social network // *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, USA, 2003. – PP. 137–146.

23. Carnes T., Nagarajan R., Wild S.M., A. Van Zuylen. Maximizing influence in a competitive social network: a follower's perspective // *Proceedings of the ninth international conference on Electronic commerce*, Minneapolis, USA, 2007. – PP. 351–360.

24. Dodds P.S., Watts D.J. A generalized model of social and biological contagion // *Journal of theoretical biology*. – 2005. – Iss. 4. – PP. 587–604.

25. Paklin N.B., Oreshkov V.I. *Biznes-analitica: ot dannyh k znaniyam*. [Business-analytics: From data to knowledge]. — SPb.: Piter, 2009. — 624 p. [in Russian]

26. *30 millions of Ukrainian accounts are registered in social networks* [Online resource]. – Mode of access: http://ain.ua/2012/09/21/96039 – Scr. title. [in Russian]

27. Sidorenko V.N., Slabchenko O.O., Morvanyuk V.A. Internet users' segmentation of the university's department based on a priory and a posteriory information in tasks of online users' activity monitoring // *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*. – Iss. 5/2011(70). – Kremenchuk, 2011. – PP. 52–54. [in Ukrainian]

28. *Description of Deductor platform* [Online resource]. – Mode of access: http://www.basegroup.ru/deductor/description – Scr. title. [in Russian]

29. Martin S., Brown W.M., Klavans R., Boyack K.W. OpenOrd: an open-source toolbox for large graph layout // *Proceedings of the SPIE 7868, Visualization and Data Analysis*, San Francisco, California, USA, 2011.

30. Force Directed Layouts [Online resource]. – Mode of access: http://philogb.github.com/blog/2009/09/30/force-directed-layouts – Scr. title.

31. Single-Link, Complete-Link & Average-Link Clustering Layouts [Online resource]. – Mode of access: http://nlp.stanford.edu/IR-book/completelink.html – Scr. title.

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

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

34. Freeman L.C. Centrality in social networks: conceptual clarification // *Social Networks*. – 1979. – Iss. 1. – PP. 215–239.

Стаття надійшла 18.01.2013.