УДК 004.89

АНАЛИЗ ДАННЫХ (DATA MINING) ОНЛАЙН СОЦИАЛЬНЫХ СЕТЕЙ С ПОМОЩЬЮ БИКЛАСТЕРИЗАЦИИ И ТРИКЛАСТЕРИЗАЦИИ

Д.В. Гнатышак
Д.И. Игнатов (dignatov@hse.ru)
С.О. Кузнецов
Й. Пульманс
А.В. Семенов
Национальный исследовательский институт Высшая школа экономики, Москва

В этой работе предлагается новый метод для анализа данных онлайновых социальных сетей. В частности, проанализированы данные сайта Вконтакте. Используя бикластеризацию выявлены группы пользователей со схожими интересами и сообщества пользователей принадлежащих сходным группам. Предпринята попытка использовать интересы пользователей как теги, для того чтобы описывать группы сети Вконтакте.

Введение

В настоящее время фокус исследований в области анализа социальных сетей сместился с одномодальных сетей, таких как сети “друг-друг”, к двумодальным [Latapy et al., 2008; Opsahl, 2011], трехмодальным [Jaschke et al., 2006; Murata, 2010; Ignatov et al., 2011] и даже многомодальным динамическим сетям [Roth et al., 2010; Yavorsky, 2011].

Это вызвано не только академическим интересом, но также современными потребностями бизнеса. Например, каждый пользователь сайта социальной сети имеет не только свой круг друзей, но и собственный профиль, например, он может принадлежать к некоторой группе пользователей, указывать свои вкусы или книги, которые он прочитал и т.д. Эти признаки профиля позволяют описать вкусы пользователя, предпочтения, пристрастия, которые релеванты для бизнес-ориентированных социальных сетей. Нахождение бисообществ

(пользователи-вкусы) или трисообществ (пользователи-предметы-теги) может помочь владельцам социальных сетей проанализировать крупные группы их пользователей и, следовательно, лучше приспособить сервис к нуждам самих пользователей, что может принести различные выгоды.

Большое количество данных социальных сетей может быть представлено двудольными и трехдольными графами. Стандартные подходы, такие как “поиск максимальных биклик” дают довольно много паттернов для анализа (в худшем случае экспоненциальное количество от размера входа). Следовательно, необходимы некоторые ослабления понятия биклики для анализа бисообществ (biclique communities).

Прикладная теория решеток предоставляет нам средства формальных понятий [Ganter et al., 1999], которые соответствуют бикликам; это хорошо известный факт в сообществе анализа данных социальных сетей ([Freeman, 1996; Duquenne, 1996; Roth et al., 2006]).

Понятийная бикластеризация [Игнатов и др., 2010] работает с масштабируемым приближением формального понятия. Перечислим достоинства понятийной бикластеризации: 1) меньшее количество паттернов для анализа; 2) меньшее время вычислений (полиномиальное против экспоненциального); 3) ручная настройка порога плотности для бикластеров (бисообществ); 4) толерантность к пропущенным парам (объект, признак).

Для анализа тримодальных сетей, таких как *фолксономии* [Wal, 2007], мы предложили метод трикластеризации [Ignatov et al., 2011]. В этой работе мы описываем новый подход – псевдотрикластеризацию для тегирования групп пользователей посредством их интересов. Этот подход отличается от традиционных методов трикластеризации, потому что он основан на извлечении бикластеров из двух отдельных объектно-признаковых таблиц. Бикластеры, которые сходны в соответствии с их *объемами*, объединяются посредством пересечения их объемов. *Содержание* первого бикластера и содержание второго бикластера становятся содержанием и модусом нового трикластера. Наш подход прошел экспериментальную проверку на выборке пользователей, их групп и интересов в социальной сети Вконтакте (http://vk.com).

1. Основные определения

Формальным контекстом в АФП [Ganter et al., 1999] называется тройка $\mathbb{K} = (G, M, I)$, где G – *множество объектов*, M – множество *признаков*, и отношение инцидентности $I \subseteq G \times M$ показывает, каким признаком обладает конкретный объект. Для любого подмножества объектов $A \subseteq G$ и подмножества признаков $B \subseteq M$ можно определить *операторы Галуа*:

$$A' = \{m \in M \mid gIm \text{ для всех } g \in A\}, \quad (1)$$
$$B' = \{g \in G \mid gIm \text{ для всех } m \in B\}.$$

Оператор " (двойное применение оператора ') образует *оператор замыкания*: идемпотентный ($A''' = A''$), монотонный ($A \subseteq B$ влечет $A'' \subseteq B''$) и экстенсивный ($A \subseteq A''$). Множество объектов $A \subseteq G$, таких что $A'' = A$, называется замкнутым. Аналогичное верно для замкнутых множеств признаков, подмножеств множества M . Пара (A, B) , такая что $A \subset G$, $B \subset M$, $A' = B$ и $B' = A$, называется *формальным понятием* контекста $\mathbb{K}$. Множества A и B замкнуты и называются *объемом* и *содержанием* формального понятия (A, B) соответственно. Для множества объектов A множество их общих признаков A' описывает сходство объектов множества A , и замкнутое множество A'' образует максимальный кластер сходных объектов (в соответствии с их общими признаками A'). Отношение “быть более общим понятием” определяется следующим образом: $(A, B) \geq (C, D)$ тогда и только тогда, когда $A \subseteq C$. Понятия формального контекста $\mathbb{K} = (G, M, I)$ упорядочены по отношению вложения объемов и образуют полную решетку, которая называется *решеткой понятий*. Для ее визуализации используются *линиевые диаграммы* (диаграммы Хассе), т.е. граф покрытия отношения “быть более общим понятием”.

Для эффективного использования АФП в случае больших контекстов часто применяют различные способы снижения числа формальных понятий (отбор понятий по размеру их объема или содержания). Альтернативный подход заключается в ослаблении определения формального понятия как максимального прямоугольника в объектно-признаковой матрице (формальном контексте). Одним из таких ослаблений является *ОП-бикластер* [Ignatov et al., 2010].

Если $(g, m) \in I$, то (m', g') называется объектно-признаковым бикластером (ОП-бикластером) с плотностью $\rho(m', g') = | I \cap (m' \times g') | /(| m' | \cdot | g' |)$.

Пусть $(A, B) \subseteq 2^G \times 2^M$ – ОП-бикластер и ρ_{min} – неотрицательное целое число, такое что $0 \leq \rho_{\min} \leq 1$, то (A, B) называется *плотным*, если он удовлетворяет ограничению $\rho(A, B) \geq \rho_{\min}$.

Приведенное определение показывает, что ОП-бикластеры отличаются от формальных понятий тем, что они не обязательно имеют единичную плотность (не все пары бикластера принадлежат *I*).

2. Описание модели и алгоритма

Пусть $\mathbb{K}_{UI} = (U, I, X \subseteq U \times I)$ – формальный контекст, который описывает интересы $i \in I$ некоторого пользователя $u \in U$. Аналогично, $\mathbb{K}_{UG} = (U, G, Y \subseteq U \times G)$ – формальный контекст, который показывает какой группе $g \in G$ принадлежит пользователь $u \in U$.

Мы можем искать плотные бикластеры как пары множеств $(users, interesets)$ в $\mathbb{K}_{UI}$, используя алгоритм ОП-бикластеризации, который описан в [Ignatov et al., 2010]. Эти бикластеры представляют собой группы пользователей, которые имеют сходные интересы. Таким же образом мы можем найти сообщества пользователей, которые принадлежат сходным группам в социальной сети Вконтакте, в виде плотных бикластеров $(users, groups)$.

Благодаря трикластеризации мы можем интерпретировать указанные интересы пользователей как теги, которые описывают похожие группы сервиса Вконтакте. Таким образом, мы решаем задачу тегирования пользователями (social tagging) и можем рекомендовать конкретному пользователю релевантные группы для вступления или интересы для отображения на личной странице, или новых друзей из интересных групп с похожими вкусами.

Для этой цели нам необходимо исследовать (формальный) триконтекст $\mathbb{K}_{UIG} = (U, I, G, Z \subseteq U \times I \times G)$, где (u, i, g) лежит в Z тогда и только тогда, когда $(u, i) \in X$ и $(u, g) \in Y$. Трикластер имеет вид $T_k = (i^X \cap g^Y, u^X, u^Y)$ для любой тройки $(u, g, i) \in Z$ с $\left|i^X \cap g^Y\right| / \left|i^X \cup g^Y\right| \geq \Theta$, где Θ заданный порог между 0 и 1. Мы можем вычислить плотность T_k непосредственным подсчетом, но это занимает время $O(|U||I||G|)$ в худшем случае, поэтому мы предпочитаем измерять качество такого трикластера по плотности бикластеров (g^Y, u^Y) и (i^X, u^X). Мы предлагаем вычислять эту оценку как

$\hat{\rho}(T_k) = \frac{\rho(g^Y, u^Y) + \rho(i^X, u^X)}{2}$; очевидно, что $0 \le \hat{\rho} \le 1$. Стоит отметить, что третья компонента (псевдо)трикластера или триадического формального понятия обычно называется *модус*.

3. Данные

Для наших экспериментов мы собрали данные из Россйиской студенческой социальной сети Вконтакте. Каждая рассматриваемая запись содержит следующие поля: id, userid, gender (пол), family status (семейный статус), birthdate (дата рождения), country (страна), city (город), institute (ВУЗ), interests (интересы), group (группа). Из этого множества мы отобрали четыре выборки по значению поля ВУЗ, а именно студентов двух ведущих технических ВУЗов и двух университетов гуманитарной и социологической направленности: Бауманский государственный технический университет (БГТУ), Московский физико-технический институт (МФТИ), Российский государственный гуманитарный университет (РГГУ) и Российский государственный социальный университет (РГСУ). Затем два формальных контекста, пользователи-интересы и пользователи-группы, были созданы для каждого из упомянутых выше наборов данных.

Таблица 1. Описание четырех наборов данных о пользователях Вконакте

	БГТУ	МФТИ	РГГУ	РГСУ
число пользователей	18542	4786	10266	12281
количество интересов	8118	2593	5892	3733
число групп	153985	46312	95619	102046

4. Эксперименты

Эксперименты проводились на ПК с процессором Intel Core i7-2600 тактовой частотой 3.4 GHz и 8 GB оперативной памяти.

Для каждого из подготовленных наборов данных были проведены следующие эксперименты. Сначала два множества бикластеров с различными ограничениями на минимальную плотность были получены для каждого формального контекста. Затем были сформированы множества бикластеров с порогом 0,5 на минимальную плотность и пары с существенным размером пересечения объема μ были отобраны и добавлены в соответствующее множество псевдо-трикластеров.

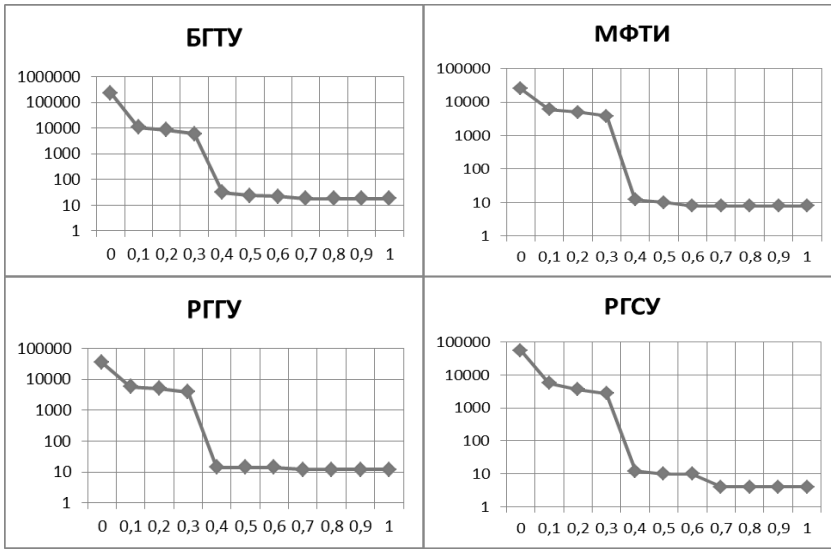

Рис. 2. Распределение плотности бикластеров для четырех Российских ВУЗов.

Как видно по графикам и результатам вычислений большинство псевдотрикластеров принимают значения $\mu \approx 0.33$. В этой серии экспериментов мы не выявили каких-либо особых интересов присущих пользователям конкретного университета, но количество бикластеров или псевдотрикластеров относительно выше для БГТУ. Это прямое следствие большего количества пользователей и разнородности их групп.

Приведем примеры плотных бикластеров и трикластеров.

Пример 1. Бикластеры в форме (Пользователи, Интересы).

$\rho = 83,33\%$, порождающая пара: $\{3609, home\}$,

бикластер: $(\{3609, 4566\}, \{family, work, home\})$

$\rho = 83,33\%$, порождающая пара: $\{30568, orthodox\ church\}$,

бикластер: $(\{25092, 30568\}, \{music, monastery, orthodox\ church\})$

$\rho = 100\%$, порождающая пара: $\{4220, beauty\}$,

бикластер: $(\{1269, 4220, 5337, 20787\}, \{love, beauty\})$

Например, второй бикластер может быть прочитан как пользователи 25092 и 30568 имеют почти все интересы “music”, “monastery”, “orthodox

church" в качестве общих. Пара генератор показывает, какая пара $(user, interest)$ была использована при построении бикластера.

Пример 2. Псевдотрикластеры $(Users, Intersts, Groups)$.

Сходство бикластеров $\mu = 100\%$, средняя плотность $\hat{\rho} = 54,92\%$.

Пользователи: $\{16313, 24835\}$,

Интересы: $\{16313, 24835\}$,

Группы: $\{365, 457, 624, \ldots, 17357688, 17365092\}$

Этот трикластер может быть интерпретирован как множество двух пользователей, которые в среднем имеют 55% общих интересов и групп. Два соответствующих бикластера имеют те же самые объемы, т.е. люди с почти всеми интересами из содержания этого трикластера и почти всеми группами из модуса совпадают.

5. Заключение

Данный подход нуждается в дальнейшем улучшении масштабируемости и качества найденных трикластеров. Мы рассматриваем несколько возможностей для приближенных вычислений; выбор хороших порогов для плотности *n*-кластеров и сходства сообществ, более сложные меры качества, такие как точность и полнота в информационном поиске. Предложенный подход нуждается в сравнении с подходами на основе решеток-айсбергов, устойчивых понятий, шумоустойчивых понятий [Besson et al, 2006] и различных подходов *n*-кластеризации из биоинформатики [Zhao et al., 2005]. Мы также утверждаем, что возможно получить более плотные псевдотрикластеры на основе традиционного анализа формальных понятий (не смотря на то, что это затратно с вычислительной точки зрения). Чтобы оценить релевантность найденных трисообществ необходима также проверка экспертом (например, социологом).

В заключении мы приходим к выводу, что возможно использовать методы псевдотрикластеризации для тегирования групп интересами пользователей на сайтах социальных сетей и находить трисообщества. Например, если мы нашли плотные трикластеры $(Users, Groups, Interests)$ мы можем пометить множество групп *Groups* интересами пользователей из множества *Interests*. Также имеет смысл применение бикластеров и трикластеров для формирования рекомендаций. Пропущенные пары и тройки являются хорошими кандидатами для рекомендаций целевому пользователю других потенциально интересных пользователей, групп и ресурсов.

Благодарности. Мы благодарим наших коллег – Винсента Дюкена, Сергея Объедкова, Камия Рота и Леонида Жукова за их вдохновляющие обсуждения, которые явно или неявно повлияли на это исследование.

Список литературы

[Besson et al, 2006] Besson, J., Robardet, C., Boulicaut, J.F.: Mining a new fault-tolerant pattern type as an alternative to formal concept discovery. LNCS Vol. 4068. Springer (2006) 144-157

[Duquenne, 1996] Duquenne, V.: Lattice analysis and the representation of handicap associations. Social Networks **18**(3) (1996) 217 – 230

[Freeman, 1996] Freeman, L.C.: Cliques, galois lattices, and the structure of human social groups. Social Networks **18** (1996) 173-187

[Ganter et al., 1999] Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. 1st edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1999)

[Ignatov et al., 2011] Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A., Zhukov, L.E.: From Triconcepts to Triclusters. In: Proceedings of the 13th international conference on Rough sets, fuzzy sets, data mining and granular computing. RSFDGrC'11, Springer (2011) 257-264

[Jäschke et al., 2006] Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: TRIAS_An Algorithm for Mining Iceberg Tri-Lattices. In: ICDM '06, Washington, DC, USA, IEEE Computer Society (2006) 907-911

[Latapy et al., 2008] Latapy, M., Magnien, C., Vecchio, N.D.: Basic notions for the analysis of large two-mode networks. Social Networks **30**(1) (2008) 31 – 48

[Murata, 2010] Murata, T.: Detecting communities from tripartite networks. In Rappa, M., Jones, P., Freire, J., Chakrabarti, S., eds.: WWW, ACM (2010) 1159-1160

[Opsahl, 2011] Opsahl, T.: Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Social Networks **34** (2011) – (in press).

[Roth et al., 2006] Roth, C., Obiedkov, S.A., Kourie, D.G.: Towards concise representation for taxonomies of epistemic communities. In Yahia, S.B., Nguifo, E.M., Belohlavek, R., eds.: CLA. LNCS Vol 4923, Springer (2006) 240-255

[Roth et al., 2010] Roth, C., Cointet, J.P.: Social and semantic coevolution in knowledge networks. Social Networks **32** (2010) 16-29

[Wal, 2007] Vander Wal, T.: Folksonomy Coinage and Definition (2007) http://vanderwal.net/folksonomy.

[Yavorsky, 2011] Yavorsky, R.: Research Challenges of Dynamic Socio-Semantic Networks. In CEUR Workshop proceedings Vol-757, CDUD'11 - Concept Discovery in Unstructured Data. (2011) 119-122

[Zhao et al., 2005] Zhao, L., Zaki, M.J.: Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data. In: SIGMOD '05, New York, NY, USA, ACM (2005) 694-705

[Игнатов и др., 2010] Игнатов Д.И., Каминская А.Ю., Кузнецов С.О., Магизов Р. А. Метод бикластеризации на основе объектных и признаковых замыканий// Интеллектуализация обработки информации: 8-я международная конференция. Сборник докладов.– М.: МАКС Пресс, 2010. – С. 140 – 143.