Метод спектральной трикластеризации для систем совместного пользования ресурсами

З. Р. Секинаева [1], Д. И. Игнатов[2]

[1]zari6a@gmail.com, [2]dignatov@hse.ru

[1,2]НИУ ВШЭ, Россия, 101000, г. Москва, ул. Мясницкая, д. 20

Аннотация. Статья посвящена разработке метода трикластеризации на основе графовой спектральной кластеризации. В серии экспериментов на реальных данных исследована эффективность и пригодность метода к анализу данных систем совместного пользования ресурсами, т.н. фолксономий.

Ключевые слова: Анализ сетевых (графовых) данных, системы совместного пользования ресурсами, спектральная трикластеризация, разработка данных (Data Mining).

Введение

За последние два года количество пользователей социальных систем использования тэгов (Social resource tagging system) таких как, Flickr (http://www.flickr.com/) – фотогалерея, del.icio.us (http://delicious.com/) – сервис закладок, Bibsonomy (http://www.bibsonomy.org/) – сервис библиографических закладок, сильно возросло. Популярность таких сервисов связана с тем, что пользователь «ставит закладку» на тот или иной ресурс, которая хранится в виде записи на удаленном сервере, и доступна с любого компьютера. Сервис социальных закладок, обладающий рекомендательным механизмом, становится, более полезным и приобретает статус рекомендательной системы на основе тегов.

Игнатов Д.И., Яворский Р.Э. (ред.): Анализ Изображений Сетей и Текстов, Екатеринбург, 16-18 марта, 2012
© Открытые системы, 2012

Трехмерная структура данных «пользователь-тэг-ресурс», так называемая фолксономия, лежит в основе таких систем. Фолксономия состоит из трёх множеств U, T, R – пользователей, ресурсов и тегов, а также тернарного отношения Y. Для поиска групп пользователей, использующих одинаковые теги для пометки некоторого множества ресурсов, применяются модели и методы анализа формальных понятий (АФП).

Анализ формальных понятий – прикладная алгебраическая дисциплина, которая развивается в течение последних пятнадцати лет и использует алгебраическую теорию решеток для формализации понятия как единицы человеческого мышления – понятия. Это означает, что понятие должно быть однородным и замкнутым. АФП опирается на так называемый диадический подход, т.е. объектно-признаковое представление данных, что дает возможность формализовать понятие как пару (объем, содержание), такой что объем – объекты, обладающие всеми признаками из содержания, а содержание – признаки, которыми обладают объекты из содержания. Хотя классическая диадическая модель успешно применяется во многих приложениях, во многих ситуациях требуется расширение размера формальных понятий, т.е. добавления третьей компоненты - условия. В качестве формального «условия» можно задавать отношения, методы, значения, причины, цели и т.д. Триадический подход в АФП основан на формализации тернарного отношения, связывающего объекты, признаки и условия [1]. Также как и диадический контекст задается с помощью двумерной таблицей, триадический контекст или триконтекст задается трехмерной.

С ростом популярности и размеров социальных систем использования тэгов, в основе которых лежит фолксономия, возникла острая необходимость в эффективных алгоритмах трикластеризации трехмерных бинарных данных, т.е. поиск плотных троек вида $\langle пользователь, тег, ресурс \rangle$. Существуют эффективные алгоритмы поиска формальных понятий и трипонятий. Но методы поиска трипонятий для реальных данных (несколько миллионов записей $\langle пользователь, тег, ресурс \rangle$), оказываются вычислительно сложными.

Кластеризация – хорошо изученная область в анализе данных. Термин «бикластеризация» впервые предложил Б.Г. Миркин в 1996 г., потому возникновение трикластеризации и n-трикластеризации было лишь делом времени. Существуют разнообразные алгоритмы бикластеризации. Авторы [2] предложили подход по поиску бикластеров, состоящих из двух множеств – объектов и признаков, а также предложили определение бикластера, который обладает полезным свойством: любое формальное понятие исходного контекста содержится в некотором

бикластере этого контекста (покомпонентное вложение) при нулевом значении порога плотности. Для случая триконтекстов авторы [3] ввели определение (плотного трикластера) и предложили алгоритм TRICL для поиска таких плотных трикластеров. Алгоритм TRIAS – один из хорошо известных алгоритмов для разработки данных фолксономий [4], который порождает формальные трипонятия.

Также существует метод для разработки данных многомерных отношений. Его реализация (Data-Peeler) [5] ищет замкнутые множества, удовлетворяющие заданным антимонотонным ограничениям, и превосходит аналогичный алгоритм CubeMiner [6] для разработки замкнутых трехмерных множеств или тримножеств.

Спектральная кластеризация может быть отнесена к дивизимной кластеризации, основанный на спектральных свойствах матрицы данных. Если рассматривать данные в виде графа, то задача кластеризации данных эквивалента задаче разбиения графа. Для двумерных данных граф является двудольным [7], для трехмерных – трехдольным [8,9].

В данной работе предлагается рассмотреть трикластеризацию трехмерных бинарных данных с помощью алгоритма спектральной трикластеризации и провести эксперименты на данных социальной системы использования тэгов Bibsonomy. Сравнить полученные данные с результатами работы алгоритмов TRICL и TRIAS.

Математическая модель

Формальный триадический контекст или *триконтекст* - это четверка (G, M, B, Y), где G – множество объектов, M - множество признаков, B – множество условий и Y – тернарное отношение между этими множествами, т.е. $Y \subseteq G \times M \times B$.

Понятием триадического контекста или *трипонятием* называется тройка (объекты, признаки, условия), которая также обладает свойствами однородности и замкнутости. Однородность достигается в случае, если каждый объект обладает каждым признаком при каждом условии в рамках трипонятия. Замкнутость достигается, когда понятие максимально относительно покомпонентного вложения.

Трикластером назовем тройку вида $T=(g^{\square}, m^{\square}, b^{\square})$, где $(g, m, b) \in Y$, где введены некоторые порождающие *бокс-операторы* (см. [3,8,9]). Плотным трикластером назовем трикластер T, плотность которого превышает минимальное значение – порог, заданный пользователем, т.е.

$$\rho(T)= |g^{\square} \times m^{\square} \times b^{\square} \cap Y| / |g^{\square} \times m^{\square} \times b^{\square}| \geq \rho_{min}.$$

Рассмотрим метод спектральной трикластеризации на примере трехмерных данных, так называемой фолксономии «пользователь-тэг-ресурс». Тернарное отношение такой структуры можно задать с помощью матрицы смежности трехдольного графа.

Матрица трехдольного графа может быть представлена в виде двумерной матрицы:

$$M = \begin{pmatrix} 0 & A_{UT} & A_{UR} \\ A_{UT}^T & 0 & A_{TR} \\ A_{UR}^T & A_{TR}^T & 0 \end{pmatrix}, \text{ где}$$

A_{UT} - матрица смежности, показывающая какие тэги какой пользователь использует. Аналогично, A_{UR} - матрица смежности пользователь-ресурс, A_{TR} - матрица смежности тэг-ресурс. Таким образом, отношение «пользователь-тэг-ресурс» в графе имеет вид треугольника, соединяющего вершины «пользователь», «тэг», «ресурс».

Пусть n – количество пользователей, m - количество тэгов, k – количество ресурсов, тогда матрица M имеет размер (n+m+k)×(n+m+k).

Разбиение трехдольного графа дает вектор собственных значений матрицы Лапласа – L, соответствующий второму наименьшему собственному значению системы обобщенной задачи нахождения собственных значений матрицы L:

$$Lx = \lambda Wx$$

Как и в случае двудольного графа, система имеет вид,

$$\begin{pmatrix} D_U & -A_{UT} & -A_{UR} \\ -A_{UT}^T & D_T & -A_{TR} \\ -A_{UR}^T & -A_{TR}^T & D_R \end{pmatrix} \begin{pmatrix} u \\ t \\ r \end{pmatrix} = \lambda \begin{pmatrix} D_U & 0 & 0 \\ 0 & D_T & 0 \\ 0 & 0 & D_R \end{pmatrix} \begin{pmatrix} u \\ t \\ r \end{pmatrix}$$

Вычислительную сложность работы алгоритма в худшем случае составляет $O((n+m+k)^3 \log(n+m+k))$.

Преобразовав систему в эквивалентную ей, получим:

$$\sigma = (1-\lambda)$$
$$A_{nUT} = (x\sigma - A_{nUR}z)y^T$$
$$A_{nUR} = (x\sigma - A_{nUT}y)z^T$$
$$A_{nTR} = (y\sigma - A_{nUT}^T x)z^T$$

Исследование этой системы с целью дальнейшей оптимизации алгоритма, а также изучение ее связей с сингулярным разложением матрицы, является предметом дальнейшей работы.

Рассмотрим пример: ученые (Jäschke, Stumme, Poelmans, Ignatov, Dedene) помечают тэгами (‘Machine Learning’, ‘Ontology’, ‘Domestic Violence’, ‘Formal Concept’, ‘Triclustering’) статьи (paper1, paper2, paper3).

Трехдольный граф отношения представлен на рисунке 1:

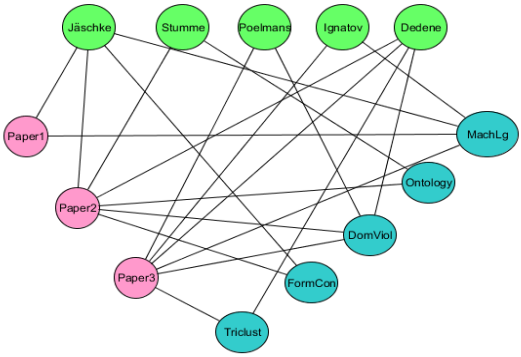


рис. 1. Пример фолксономии (статьи, читатели, теги)

Рекурсивный вызов функции кластеризации или разбиения графа дает следующую иерархию трикластеров. Критерий остановки – трикластер, содержащий либо одного пользователя, либо один тэг, либо один ресурс.

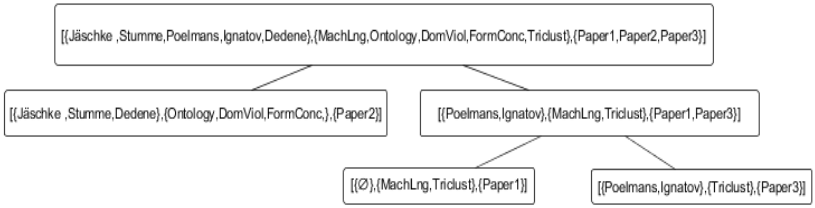


рис. 2. Пример дерева работы алгоритма SpecTric

Итого имеем 3 кластера (исходная матрица не рассматривается в качестве трикластера). Для интерпретации данных интересны лишь те трикластеры, которые содержат более одного пользователя, тэга и ресурса.

Эксперименты на данных системы социальных закладок: Bibsonomy.org

Для экспериментов были выбраны реальные данные сервиса Bibsonomy.org, содержащие 816197 записей вида (пользователь, тег, ресурс). Фактически мы имеем триконтекст, которому соответствует параллелепипед размерами |U||T||R|=2337×67464×28920, с принадлежа-

щими ему $|Y| = 816\,197$ тройками. В качестве экспериментального оборудования использовался компьютер с процессором Intel(R) Core(TM) Duo CPU с частотой 2,1ГГц и с 4ГБ оперативной памяти. В качестве языка программирования выбран высокоуровневый язык Matlab версии 7.8.0(R2009a). Предварительно данные были обработаны: все тэги были приведены к нижнему регистру.

Эксперименты проводились на 100 тысячах исходных записей, которые соответствуют параллелепипеду размером $59 \times 5197 \times 28920$.

На графике представлена структура матрицы смежности графа для первых ста тысяч троек данных сервиса Bibsonomy.org

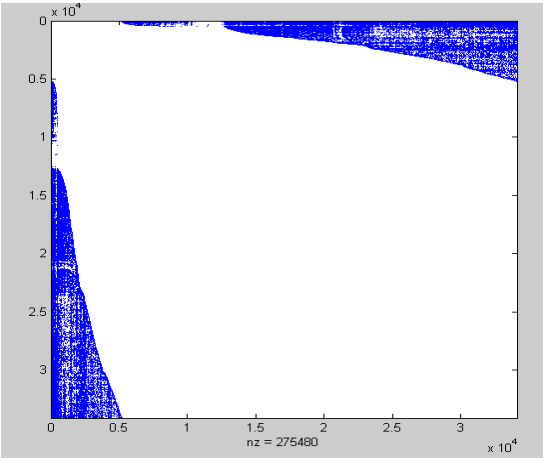


рис. 3. Вид исходной матрицы M с помощью команды spy в системе Matlab

Плотность матрицы: $\rho = \frac{|Y|}{|U||T||R|} = 0{,}000113.$

Перед применением алгоритма интересно посмотреть статистическое распределении. На рисунке 1 показаны гистограммы распределения тэгов по количеству внесенных в систему пар («пользователь», «документ») для первых ста тысяч троек. Данные следуют степенному закону распределения $p(x) = Cx^{-\alpha}$, $\alpha = 3{,}6778$ и $\sigma = 0{,}0001$ в случае документов и количества пар (пользователь, тэг). Для пользователей и тегов $\alpha = 2{,}13$ и $\alpha = 1{,}8$ соответственно. Это наблюдение позволяет применять жадную стратегию поиска, то есть находить большие и плотные трикластеры, так как небольшая часть пользователей совершает присваивание (тэг, документ). Аналогично, для тэгов и документов.

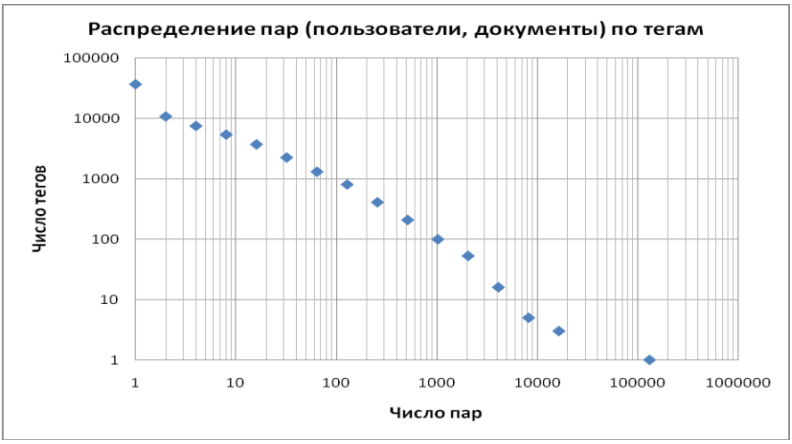


рис. 4. Распределение числа пар (пользователи, документы) по тегам

Для первых ста тысяч записей алгоритм спектральной трикластеризации породил 131 трикластер приблизительно за 9 минут 14 секунды (без учета подсчета плотностей трикластеров) в то время как алгоритм TRICL породил 4462 трикластера примерно за 21 минуту (без учета подсчета плотностей трикластеров).

Время работы алгоритмов увеличивается при подсчете плотностей, т.к. при этом осуществляется полная проверка на принадлежность исходному отношению I тройки (g, m, b) – элемента конкретного трикластера T.

В таблице представлены результаты экспериментов на первых ста, десяти тысяч, ста и двухстах тысяч троек данных сервиса Bibsonomy с подсчетом плотности трикластеров (ρ_{min}=0):

табл. 1. Результаты работы работы алгоритмов SPECTRIC и TRICL

Кол-во записей	|U|	|T|	|R|	$\|T_c\|$	$\|T_s\|$	TRICL, s	SPECTRIC, s
100	1	47	52	1	1	0.2	0.2
10 000	1	395	5193	1	1	2	2
100 000	59	5138	28920	4462	131	10311	6215

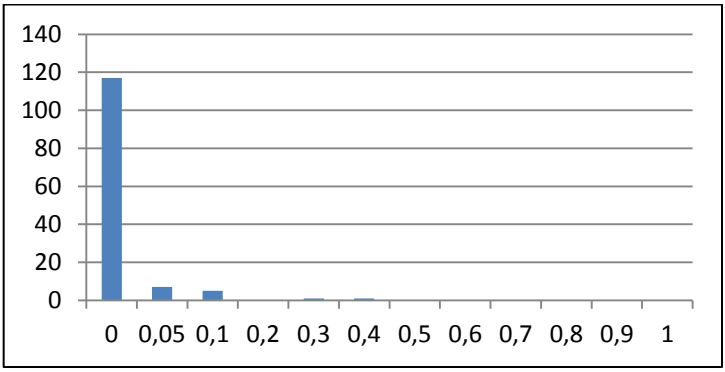


рис. 5. Гистограмма плотностей трикластеров для первых ста тысяч записей:

Также дополнительные эксперименты показали, что метод является устойчивым при 15% уровне шума (имеется в виду инвертирование тройки (g, m, b) в соответствии с равномерным распределением), но перестает корректно выявлять подгруппы при уровне шума 20% и более.

Основная доля трикластеров, порожденных в ходе эксперимента, имеет довольно невысокие значения плотности.

Заключение

Нахождение трикластеров позволяет видеть все ресурсы (их идентификаторы) и тэги, которыми пользовались клиенты сервиса Bibsonomy.

Результаты, полученные в ходе экспериментов позволяют сделать следующие выводы: на реальных данных алгоритм спектральной трикластеризации работает в два раза быстрее существующих алгоритмов, количество порождаемых трикластеров в среднем в 30 раз меньше по сравнению с количеством трикластеров, представляющими собой формальные понятия (алгоритм TRIAS), и количеством трикластеров порождаемых алгоритмом TRICL, предложенным в работах [3,8].

Однако данные трикластеры не такие плотные по сравнению с трикластерами, порождаемыми алгоритмом TRICL, и трикластерами, представляющими собой формальные понятия, плотность которых всегда равна единице.

Эксперименты на синтетических данных позволяют говорить, что при 15% уровне «шума», то есть разного рода интересах участников сообщества, алгоритм корректно выделяет сообщества (алгоритм перестает корректно выявлять подгруппы при уровне шума 20%).

Дальнейшие исследования по данной тематике возможны в двух направлениях: с одной стороны это улучшение эффективности алгоритма трикластеризации с использованием аппарата линейной алгебры, а именно, исследования связи полученной в работе системы с сингулярными значениями матрицы, с другой стороны, введение дополнительных критериев остановки ветвления алгоритма и разработки новых мер качества.

Список источников

1 Lehmann F., Wille R.: A triadic approach to formal concept analysis. In: Ellis, G., Rich, W., Levinson, R., Sowa, J.F. (eds.) ICCS 1995. LNCS, vol. 954, pp. 32–43 Springer, Heidelberg, 1995.

2 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.

3 Игнатов Д.И., Магизов Р.А. Анализ тримодальных данных на примере Интернет-сервисов социальных закладок// Социологические методы в современной исследовательской практике: Сборник статей. – М.: НИУ ВШЭ, 2011. – С. 315 – 321.

4 Robert Jäschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, Gerd Stumme: TRIAS - An Algorithm for Mining Iceberg Tri-Lattices. ICDM 2006:907-911.

5 Loïc Cerf, Jérémy Besson, Céline Robardet, Jean-François Boulicaut: Data Peeler: Contraint-Based Closed Pattern Mining in n-ary Relations. SDM 2008:37-48.

6 Liping Ji, Kian-Lee Tan, Anthony K. H. Tung: Mining Frequent Closed Cubes in 3D Datasets. VLDB 2006:811-822

7 Zhukov L. Technical report Spectral Clustering of Large Advertiser Datasets Part 1. 2004.

8 Dmitry I. Ignatov, Sergei O. Kuznetsov, Ruslan A. Magizov, Leonid E. Zhukov: From Triconcepts to Triclusters. RSFDGrC 2011, LNCS/LNAI Volume 6743, Springer, 2011:257-264.

9 Игнатов Д. И., Кузнецов С. О., Пульманс Й. Разработка данных систем совместного пользования ресурсами: от трипонятий к трикластерам //Математические методы распознавания образов: 15-я Всероссийская конференция. Сборник докладов. – М.: МАКС Пресс, 2011. – С. 258 – 261.