Рекомендательная система парфюмерной продукции и ее тегов на основе трикластеризации

Венжега А. В., Гнатышак Д. В., Игнатов Д. И., Константинов А. В.
dignatov@hse.ru
НИУ ВШЭ, Москва, Россия

В статье приводится экспериментальное сравнение трех алгоритмов трикластеризации в задаче рекомендации тегов и ресурсов. Рассматриваются реальные данные интернет-магазина SpellSmell.ru. Результаты оцениваются в терминах точности и полноты

Recommender system for perfumes and their tags based on triclustering

Venjega A. B., Gnatyshak D. V., Ignatov D. I., Konstantinov A. V.
NRU HSE, Moscow, Russia

In this paper we show the results of the comparison of three triclustering algorithms for the purpose of tags and resources recommendation. We consider the data of the online store SpellSmell.ru. We assess the results by precision and recall.

С начала 2000-ых годов задача разработки качественных рекомендательных систем стала одной из ключевых в сфере Интернет-коммерции. С тех пор было предложено множество различных подходов по построению рекомендательных систем. В связи с тернарной природой отношений некоторых входных данных возникает необходимость в рекомендательных системах особого рода. В данной работе предлагается необычный подход к созданию рекомендательных систем на основе трикластеризации данных парфюмерного онлайн-магазина `www.SpellSmell.ru`. Также приводится оценка качества полученной рекомендательной системы на основе показателей точности и полноты и производится ее сравнение с уже существующими рекомендательными системами.

Методы трикластеризации

Трикластеризация объект-признак-условие. В системе использовалось две разновидности данного метода трикластеризации: на основе бокс-операторов [3] и на основе штрих-операторов.

Для начала определим алгоритм трикластеризации объект-признак-условие (OAC-трикластеризации) на основе бокс-операторов. Пусть дан триадический контекст $\mathbb{K} = (G, M, B, I)$, где G, M, B — множества, а $I \subseteq G \times M \times B$ — тернарное отношение. Для фиксированной тройки $(\tilde{g}, \tilde{m}, \tilde{b}) \in I$ определим бокс-операторы:

$$\tilde{g}^{\square} := \{\, g \,|\, (g, m) \in \tilde{b}' \vee (g, b) \in \tilde{m}' \} \quad (1)$$

$$\tilde{m}^{\square} := \{\, m \,|\, (g, m) \in \tilde{b}' \vee (m, b) \in \tilde{g}' \} \quad (2)$$

$$\tilde{b}^{\square} := \{\, b \,|\, (g, b) \in \tilde{m}' \vee (m, b) \in \tilde{g}' \} \quad (3)$$

где $(\cdot)'$ — штрих-оператор для триадического случая [5].

Тогда назовём *OAC-трикластером на основе бокс-операторов*, построенным для тройки $(g, m, b) \in I$, тройку множеств $T = (g^{\square}, m^{\square}, b^{\square})$. Его компоненты будем называть, по аналогии с трипонятием, *объёмом*, *содержанием* и *модусом*.

Также необходимо определить плотность трикластера: $\rho(X, Y, Z) = \frac{|I \cap (X \times Y \times Z)|}{|X||Y||Z|}$.

Метод последовательно перебирает все тройки контекста и для каждой из них получает трикластер, который добавляется в общее множество трикластеров. Для отслеживания повторного порождения трикластера рекомендуется использование хеш-значений, что даёт значительный выигрыш по времени. Возможно также задание порога на минимальную плотность трикластера.

Теперь перейдём к трикластеризации объект-признак-условие на основе штрих-операторов. Данный метод использует несколько иную схему построения трикластеров, являющуюся, по сути, расширением на триадический случай метода, описанного в [9]. Пусть имеется триадический контекст $\mathbb{K} = (G, M, B, I)$. Выпишем явно штрих-операторы, используемые методом:

$$(g, m)' = \{\, b \,|\, (g, m, b) \in I \} \quad (4)$$

$$(g, b)' = \{\, m \,|\, (g, m, b) \in I \} \quad (5)$$

$$(m, b)' = \{\, g \,|\, (g, m, b) \in I \} \quad (6)$$

Тогда назовём *OAC-трикластером, основанным на штрих-операторах*, построенным на тройке $(g, m, b) \in I$, тройку множеств $T = ((m, b)', (g, b)', (g, m)')$.

Псевдокод алгоритма данного вида трикластеризации также приведён ниже.

Метод на основе анализа формальных понятий: TRIAS. Метод TRIAS, описанный в [4] является методом поиска триадических формальных понятий, которые, впрочем, можно рассматривать как абсолютно плотные трикластеры.

TRIAS основан на алгоритме NextClosure, который находит все формальные понятия диади-

Алгоритм 1. Алгоритм OAC-трикластеризации, основанной на штрих-операторах.

Вход: $\mathbb{K} = (G, M, B, I)$ — триконтекст;
ρ_{min} — порог плотности
Выход: $TSet = \{(X, Y, Z)\}$
для всех (g, m): $g \in G$,$m \in M$
  $PrOA[g, m] = (g, m)'$
для всех (g, b): $g \in G$,$b \in B$
  $PrOC[g, b] = (g, b)'$
для всех (m, b): $m \in M$,$b \in B$
  $PrAC[m, b] = (m, b)'$
для всех $(g, m, b) \in I$
  $T = (PrAC[m, b], PrOC[g, b], PrOA[g, m])$
  $Tkey = hash(T)$
  если $Tkey \notin Tset.keys \wedge \rho(T) \geqslant \rho_{min}$ **то**
    $Tset[Tkey] = T$

ческого контекста, перебирая их в лексикографическом порядке. Он расширяет данный алгоритм на триадический случай, а также добавляет условия минимальной поддержки, т.е. формальные понятия со слишком маленькими объёмом, содержанием и/или модусом будут отбрасываться.

Рекомендация тегов и ресурсов

Одной из особенностей Web 2.0 является создаваемый пользователями Интернет-контент и тесное взаимодействие между посетителями и веб-ресурсом. Следствием этого является перегруженность Интернет-ресурсов различными не структурированными данными. Одним из способов решения данной проблемы является введение возможности классифицировать и описывать данные самими пользователями путем использования тэгов. Напомним, что под термином тэг понимается свободно выбранное ключевое слово, описывающее данные или информацию. Группируя и описывая данные с помощью тэгов мы можем сделать контент ресурса более структурированным и пригодным для ориентирования и поиска нужных объектов. Чтобы помочь пользователям в процессе присвоения тэгов, большинство популярных веб-сервисов, таких как `YouTube.com` и `Flickr.com`, предлагают некоторую систему рекомендации тэгов. Результатом работы данных систем является структурированность информации на веб-сервисе, более высокая вероятность того, что на данный ресурс будут вести внешние ссылки и, как следствие, более высокое место данного сервиса в поисковых машинах и высокая привлекательность среди пользователей.

Стоит также отметить, что когда пользователь веб-ресурса присваивает наименованию определенный тэг, т. е. порождает отношение между пользователем и наименованием с помощью тэга, образуется фолксономия. Фолксономия $\mathbb{F}$ состоит из трёх множеств: U — пользователей, R — ресурсов и T —

Алгоритм 2. Алгоритм рекомендаций на основе трикластеризации

Вход: $\mathbb{F} = (U, T, R, Y)$ — фолксономия;
Выход: Res_{rec}, Tag_{rec} — рекомендации тегов и ресурсов
Получить множество трикластеров $\mathcal{T}$
для всех $u \in U$
  для $i = 1, \ldots, |\mathcal{T}|$
    $sim_u(\mathcal{T}_i) = \frac{1}{2}\left(\frac{|R_u \cap R_{\mathcal{T}_i}|}{|R_u \cup R_{\mathcal{T}_i}|} + \frac{|T_u \cap T_{\mathcal{T}_i}|}{|T_u \cup T_{\mathcal{T}_i}|}\right)$
  $\mathcal{T}_{best} = \arg\max sim_u(\mathcal{T}_i)$
  $Res_{rec}[u] = R_{\mathcal{T}_{best}} \setminus R_u$
  $Tagrec[u] = T_{\mathcal{T}_{best}} \setminus T_u$

тэгов и тернарного отношения $Y \subseteq U \times T \times R$ между ними. Эквивалентным и более наглядным представлением фолксономии является неориентированный гиперграф $G = (V, E)$, где $V = U \cup T \cup R$ — набор вершин, и $E = \{\{u, t, r\} \mid (u, t, r) \in Y\}$ — множество граней. Пользователи могут помечать ресурсы несколькими тэгами и наоборот, помечать одним тэгом несколько наименований.

Таким образом, видно, что фолксономии являются рановидностями триадического контекста, следовательно, к ним применимы алгоритмы трикластеризации. Предложенный в данной работе метод рекомендаций на основе трикластеризации состоит в том, чтобы из множества полученных для фолксономии трикластеров найти наиболее похожий на тройки с заданным пользователем и выдать его содержание и модус в качестве рекомендаций.

Выше приведён псевдокод алгоритма данного метода.

Оценка качества. Чаще всего для оценки информационного поиска используются два показателя: точность (precision) и полнота (recall). Они определены для простого случая, когда информационно-поисковая система возвращает набор документов, соответствующих запросу [10]. Рекомендательные системы схожи с методами информационного поиска и оцениваются с помощью аналогичных показателей. Для оценки предложенной в данной работе рекомендательной системы использовались показатели точности и полноты. Их значения для рекомендации ресурсов (для тэгов вычисляются аналогично) вычисляются следующим образом:

$$Precision = \frac{|Res_{test} \cap Res_{rec}|}{|Res_{rec}|} \tag{7}$$

$$Recall = \frac{|Res_{test} \cap Res_{rec}|}{|Res_{test}|} \tag{8}$$

Способ подсчёта данных показателей представлен в алгоритме 3.

Алгоритм 3. Алгоритм подсчёта точности и полноты

Вход: $\mathbb{F} = (U, R, T, Y)$ — фолксономия
s_{min} — минимальное допустимое число троек пользователя
Выход: $recall_{tag_{av}}$, $prec_{tag_{av}}$, $recall_{res_{av}}$, $prec_{res_{av}}$

$supp(u_i) = |\{(u_i, t, r) \in Y)\}|$
$u'_i = \{(t, r) \,|\, (u_i, t, r) \in Y\}$
$n := 0$
для i=1,. . . ,|U|
  если $supp(u_i) > s_{min}$ **то**
    $n := n + 1$
    разбить u'_i на четыре (почти) равномощных множества: $u'_i = \bigsqcup_{l=1}^{4} P_l$
    для $j = 1, \ldots, 4$
      $testSet := P_j$; $trainingSet := \bigcup_{l \neq j} P_l$
      найти рекомендации Res_{rec} и Tag_{rec} для u_i
      вычислить $Prec_{tag}[n][j]$, $Recall_{tag}[n][j]$, $Prec_{res}[n][j]$ и $Recall_{res}[n][j]$
$recall_{tag_{av}} := \frac{\sum_{i,j} Recall_{tag}[i][j]}{4n}$
$prec_{tag_{av}} := \frac{\sum_{i,j} Prec_{tag}[i][j]}{4n}$
$recall_{res_{av}} := \frac{\sum_{i,j} Recall_{res}[i][j]}{4n}$
$prec_{res_{av}} := \frac{\sum_{i,j} Prec_{res}[i][j]}{4n}$

После подсчёта точности и полноты, их значения агрегировались с помощью $F1$-меры:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{9}$$

Машинные эксперименты

Данные. Входными данными, использующимися всеми тремя алгоритмами трикластеризации, являются тройки, отражающие действия пользователя сайта парфюмерии `SpellSmell.ru`. Данные имеют вид таблицы, каждая строка которой является тройкой вида (номер_пользователя, номер_ресурса, номер_тэга) и состоят из 10000 таких троек. Объем входных данных следующий: $|Y| = 10000$, $|U| = 723$, $|R| = 3876$, $|T| = 19$.

Вычислив P-значение (0,648), мы можем сделать вывод о том, что наименьшая величина уровня значимости, при котором нулевая гипотеза о степенном распределении отвергается, достаточно высока, а значит, распределение входных данных подчиняется степенному закону (power law).

Данный анализ влечет за собой интересные свойства согласно известному правилу «20:80»:

$$W = P^{\frac{\alpha-2}{\alpha-1}} \tag{10}$$

В нём говорится, что доля W благосостояния находится в руках P богатейших людей населения

Таблица 1. Результаты трикластеризации

	ρ_{av}	$\lvert T\rvert$
OAC (бокс)	0,02	535
OAC (штрих)	0,8033	1606
TRIAS	1	583

Таблица 2. Значения полноты, точности и $F1$-меры для рекомендации ресурсов

Метод	Полнота	Точность	$F1$-мера
	$s_{min} = 10$		
OAC (бокс)	0,0242	0,059	0,0343
OAC (штрих)	0,0631	0,1269	0,0843
TRIAS	0,14	0,1929	0,1622
	$s_min = 60$		
OAC (бокс)	0,02	0,102	0,0334
OAC (штрих)	0,0424	0,1921	0,0695
TRIAS	0,0624	0,225	0,0977

[7]. В нашем случае ($\alpha = 2.42$) данное правило интерпретируется как «47% пользователей осуществляют 80% действий над ресурсами (помечают ресурс тэгами)». Следовательно, для рекомендаций наиболее важна имена эта часть пользователей, дающая наибольшее число оценок.

Рекомендация тегов и ресурсов.

Для алгоритмов трикластеризации экспериментальным путём были определены следующие значения параметров:

1) OAC (бокс): $\rho_{min} = 0,011$;
2) OAC (штрих): $\rho_{min} = 0,5$;
3) TRIAS: $\tau_U = 1$, $\tau_T = 2$, $\tau_R = 5$.

С учётом данных значений параметров были получены результаты, кратко представленные в таблице 1.

В качестве рекомендаций выдавались 10 первых подходящих элементов содержания или модуса трикластера. В результате проведения n-кратной перекрестной проверки были получены значения точности и полноты для различных алгоритмов трикластеризации (таблицы 2 и 3).

Проанализировав полученные результаты для ограничений $s_{min} = 10$ и $s_{min} = 60$, можно сделать вывод о том, что увеличение порогового значения количества порожденных пользователем троек не влечет за собой увеличения значения $F1$-меры, а значит не влияет на качество рекомендаций в положительную сторону.В итоге, сравнение алгоритмов производилось при ограничении $s_{min} = 10$ по значениям $F1$-меры.

Столь невысокие показатели точности, полноты и $F1$-меры для рекомендации тэгов объясняются

Таблица 3. Значения полноты, точности и $F1$-меры для рекомендации тэгов

Метод	Полнота	Точность	$F1$-мера
	$s_{min} = 10$		
OAC (бокс)	0,0679	0,3114	0,1115
OAC (штрих)	0,0679	0,3114	0,1115
TRIAS	0,0437	0,2694	0,0752
	$s_{min} = 60$		
OAC (бокс)	0,0309	0,4089	0,0575
OAC (штрих)	0,0309	0,4089	0,0575
TRIAS	0,0228	0,3666	0,0429

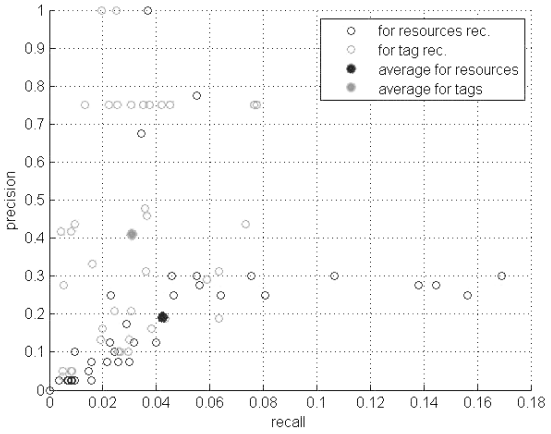

Рис. 1. Пример значений полноты и точности алгоритма OAC (штрих) при $s_{min} = 60$

низким значением мощности множества тэгов $|T| = 19$ при значительно большем числе троек $|Y| = 10000$.

Как видно из таблицы 2, наиболее качественным алгоритмом для рекомендательной системы ресурсов является TRIAS (значение $F1$-меры — 0,162), а в случае рекомендации тэгов (таблица 3) наиболее качественные результаты показали алгоритмы OAC-трикластеризации (0,112).

Данные результаты выглядят достаточно низкими по сравнению с системами извлечения информации, в которых значения точности и полноты превышают 50% порог. Однако, в случае рекомендательных систем, данные результаты являются достаточно высокими. Наример, в соревновании алгоритмов рекомендации для BibSonomy конференции ECML PKDD 2009 значение $F1$-меры победителей [6] равнялось 0.187, а команды, занявшей 4-ое место, — 0.141 (результат 16-ой команды — 0.046).

Выводы

В данной работе рассмотрены три алгоритма трикластеризации: OAC-трикластеризация на основе бокс-операторов и штрих-операторов и алгоритм TRIAS. Предложенная рекомендательная система использует необычный подход для решения проблемы построения рекомендации для триадических данных. Проведена оценка алгоритмов, использованных в данной системе, и выявлены наиболее подходящие алгоритмы для конкретных целей и рекомендаций. С помощью метода n-кратной перекрестной проверки оценено качество полученной рекомендательной системы. Проведенная оценка качества рекомендательной системе говорит об относительно высокой степени точности и полноты предоставляемых ею рекомендаций. Результатом работы прототип рекомендательной системы для триадических данных, пригодный к использованию в современных веб-сервисах.

Литература

[1] *Ganter B., Wille R.* Formal concept analysis: Mathematical Foundations — Berlin-Heidelberg: Springer, 1999.

[2] *Ricci F., Rokach L.* Recommender Systems Handbook — Springer Science + Business Media, LLC, 2011.

[3] *Ignatov D. I., Kuznetsov S. O., Magizov R. A., Zhukov L. E.* From Triconcepts to Triclusters // 13-th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC-2011), 2011. — Pp. 257-264.

[4] *Jaschke R., Hotho A., Schmitz C., Ganter B., Stumme G.* TRIAS - An Algorithm for Mining Iceberg Tri-Lattices // ICDM, 2006. — Pp. 907-911.

[5] *Lehmann F., Wille R.* A triadic approach to formal concept analysis // Conceptual Structures: Applications, Implementation and Theory. Springer, 1995. — Pp. 32-43.

[6] *Lipczak M., Hu Y., Kollet Y., Milios E.* Tag Sources for Recommendation in Collaborative Tagging Systems // ECML PKDD Discovery Challenge, 2009.

[7] *Newman M. E. J* Power Laws, pareto distributions and Zipf's law // Contemprorary Physics, 2005. — Vol. 46, No. 5. — Pp. 323-351

[8] *Гнатышак Д. В., Игнатов Д. И., Жуков Л. Е., Кузнецов С. О., Миркин Б. Г.* Экспериментальное сравнение некоторых алгоритмов трикластеризации // Международная конференция «Интеллектуализация обработки информации» (ИОИ-9), 2012. — С.??–??.

[9] *Игнатов Д. И., Каминская А. Ю., Кузнецов С., О., Магизов Р. А.* Метод бикластеризации на основе объектных и признаковых замыканий // Международная конференция «Интеллектуализация обработки информации» (ИОИ-8), 2010. — С.140–143.

[10] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск : Пер. с англ. — М. : ООО «И. Д. Вильямс», 2001