этих изменений и связей, объединение точечных мнений в единую структуру, определение соотношения социального и индивидуального), важно отметить недостатки. На данном этапе развития процедура дискурс-анализа трудоемка и длительна при обработке больших массивов информации, также нельзя полностью исключить влияние личности исследователя. Важная особенность дискурс-анализа: его использование целесообразно перед проведением количественного исследования, в условиях, когда объект будущего незнаком или изменил свои основные свойства.

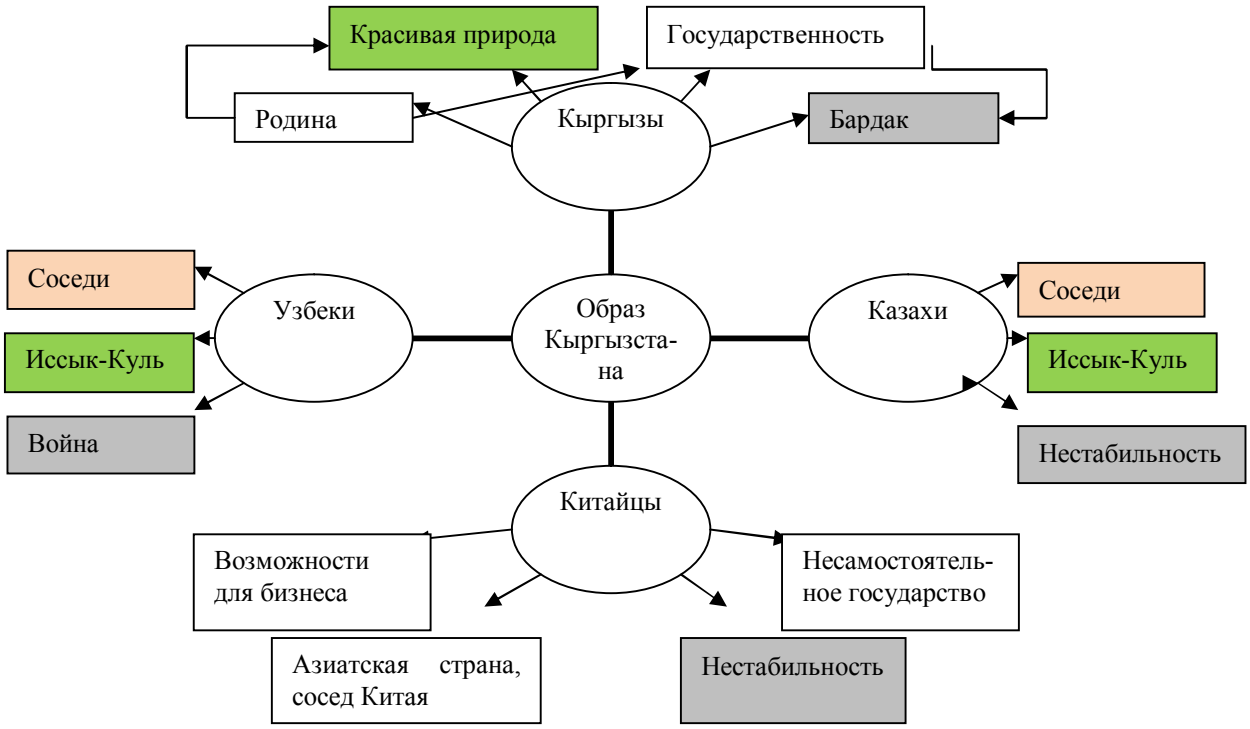

Рис. Образ Кыргызстана в восприятии кыргызов, узбеков, казахов, китайцев

В результате использования метода дискурс-анализа удалось объединить смысловые основы дискурсов отдельных людей в единую систему. В рамках подтемы интервью «Образ Кыргызстана» были определены установки информантов и их истоки, а также аргументы, способные модифицировать установки и направить поведение в определенное русло.

Анализ тримодальных данных на примере Интернет-сервисов социальных закладок

Игнатов Дмитрий Игоревич, *НИУ ВШЭ*
Магизов Руслан Азатович, *НИУ ВШЭ*

Введение

В работе представлен оригинальный метод трикластеризации, который использовался для поиска «плотных» троек вида (посетители, тэги, ресурсы) при анализе данных сервиса социальных закладок Bibsonomy. Предложенный нами метод является универсальным для анализа любых тримодальных данных, описываемых тернарными отношениями, и служит вычислительно эффективной альтернативой поиску трипонятий (в смысле анализа формальных понятий) и построению айсбергов трирешёток.

В настоящее время есть большое количество Интернет-сервисов, предлагающих совместное использование различных электронных ресурсов (social resource sharing systems), сгенерированных пользователями. Среди таких систем можно отметить Flickr (http://www.flickr.com/) — фотогалерея, del.icio.us (http://delicious.com/) — сервис закладок, Bibsonomy (http://www.bibsonomy.org/) — сервис библиографических закладок; среди российских сервисов

социальных закладок можно упомянуть Bobrdobr.ru (http://www.bobrdobr.ru/) и Яндекс.Закладки (http://zakladki.yandex.ru/). Ключевая особенность таких сервисов в том, что пользователь «ставит закладку» на тот или иной ресурс, которая хранится в виде записи на удаленном сервере, а потому доступна с любого компьютера, вне зависимости от браузера. Стоит заметить, что современные веб-браузеры могут хранить закладки на сетевые ресурсы и поддерживать их синхронизацию через Интернет (например, Opera), но они не выполняют функцию совместного пользования. Помимо закладки пользователь такой системы может присвоить размещаемому им ресурсу (фотографии, Интернет-ссылке, библиографической записи) короткие ключевые слова, или тэги. С помощью таких тэгов пользователями независимо друг от друга осуществляется коллективная классификация ресурсов сервиса. Такие сервисы называют также коллективными системами использования тэгов (*Collaborative Tagging Systems)*. Зная тэги, которые использует посетитель, можно находить пользователей, помечающих ресурсы такими же ключевыми словами, и рекомендовать такие ресурсы в качестве потенциально интересных, а самих пользователей — в качестве возможных друзей (людей с близкими вкусами). Сервис социальных закладок, снабженный таким рекомендательным механизмом, становится более полезным и приобретает статус рекомендательной системы на основе тэгов (несколько лет назад такое превращение произошло с сервисом del.icio.us, что повысило его популярность).

В основе социальных сервисов совместного использования ресурсов лежит структура данных, называемая фолксономией (folksonomy)[1]. Фолксономия состоит из трёх множеств U, T, R — пользователей, ресурсов и тэгов, а также тернарного отношения Y между ними. Таким образом, к двум измерениям традиционной модели анализа потребительской корзины (покупки (transactions) и товары (items)) мы получили дополнительное измерение. И если в задаче анализа потребительской корзины отыскиваются (крупные) группы покупателей, приобретающие одинаковые продуктовые наборы, то при анализе фолксономий существенно находить группы пользователей, использующие одинаковые тэги для пометки некоторого множества ресурсов. Владельцы сервиса Bibsonomy.org используют модели и методы анализа формальных понятий (АФП)[2] для поиска таких групп (пользователи, ресурсы, тэги). Анализ формальных понятий является прикладной алгебраической дисциплиной, которая использует алгебраическую теорию решёток для формализации основной единицы человеческого мышления — понятия. Опираясь на объектно-признаковое представление данных, можно провести строгую формализацию понятия, как пары, состоящей из объема и содержания. Объем состоит из множества всех объектов, обладающих всеми признаками из содержания, а содержание — из множества всех общих признаков этих объектов; т.е. выполняется условие замкнутости, максимальности размера объема и содержания. Для формальных понятий также справедлив знаменитый закон обратного соотношения между размером объема и содержания понятия. На множестве формальных понятий можно построить частичный порядок (иерархию) по отношению «быть более общим понятием», в котором понятие с большим по вложению объемом будет более общим для понятия с меньшим объемом. Данный частичный порядок обладает решёточными свойствами и называется решеткой понятий. Исходные объектно-признаковые данные получили название формального контекста в АФП, а для фолксономий (объекты – пользователи, признаки – тэги) исходные данные содержат одно дополнительное множество, которое в АФП принято называть условиями (для фолксономий это ресурсы). Существует так называемый триадический анализ формальных понятий[3], расширение классической диадической модели, который имеет дело с трипонятиями. Такое трипонятие по-прежнему однородно и замкнуто, т.е. представляет собой максимальный параллелепипед в исходной таблице данных вида объекты-признаки-условия. Третья компонента понятия называется модусом. Существуют эффективные алгоритмы поиска формальных по-

[1] Jäschke R., Marinho L.B., Hotho A., Schmidt-Thieme L., Stumme G. Tag Recommendations in Folksonomies // Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17–21, 2007. 2007. P. 506–514.

[2] Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. Berlin et al.: Springer, 1999.

[3] Lehmann F., Wille R. Triadic Approach to Formal Concept Analysis // Proceedings of the Third International Conference on Conceptual Structures: Applications, Implementation and Theory. Santa Cruz, 1995. P. 32–43.

нятий, на основе которых построены методы поиска трипонятий. Но в виду большой вычислительной сложности поиска формальных понятий, экспоненциальной в худшем случае от размера входа, методы поиска всех формальных трипонятий для реальных данных (несколько миллионов записей $\langle$*пользователь*, *тег*, *ресурс*$\rangle$), оказываются вычислительно неприемлемыми. Одним из традиционных решений является сокращения числа понятий, оставляемых для их дальнейшего анализа, на основе отбора по размеру объема или содержания. Таким путем пошли владельцы ресурса Bibsonomy.org. Нами ранее был предложен подход к поиску бикластеров[1], состоящих их двух множеств — объектов и признаков соответственно, — который оказался вычислительно более эффективным. Помимо этого, предложенное нами определение бикластера обладает полезным свойством: любое формальное понятие исходного контекста содержится в некотором бикластере этого контекста (в смысле покомпонентного вложения) при нулевом значении порога плотности. В этой работе мы обобщаем данный подход для случая триконтекстов (фолксономий), вводим определение (плотного) трикластера, предлагаем алгоритм их поиска и проводим эксперименты на данных сервиса Bibsonomy.org.

Математическая модель и алгоритм

Напомним определения диадического АФП. Формальным контекстом называется тройка (G, M, $I \subseteq G \times M$), где G — множество объектов, M — множество признаков, а I — отношение инцидентности, показывающее, что данный объект g обладает признаком m (записывается gIm). Операторы Галуа сопоставляют подмножеству множества объектов множество всех их общих признаков или подмножеству множества признаков множество всех объектов, которые ими обладают. Пусть $A \subseteq G$, а $B \subseteq M$, тогда операторы Галуа задаются следующими выражениями:

$$A' = \{m \mid m \in A \text{ и } (g,m) \in I\}$$
$$B' = \{g \mid g \in B \text{ и } (g,m) \in I\}.$$

Формальным понятием называется такая пара (A, B), где $A \subseteq G$, а $B \subseteq M$, что $A' = B$, а $B' = A$.

Дадим определение трипонятия триадического контекста (G, M, B, Y) как тройки вида (A_1, A_2, A_3), максимальной в смысле покомпонентного вложения, где $A_1 \subseteq G$, $A_2 \subseteq M$, $A_3 \subseteq B$, такой что для $X_1 \subseteq G$, $X_2 \subseteq M$, $X_3 \subseteq B$ и $X_1 \times X_2 \times X_3 \subseteq Y$, из условия $A_1 \subseteq G$, $A_2 \subseteq M$, $A_3 \subseteq B$ всегда следует (A_1, A_2, A_3)= (X_1, X_2, X_3). Как и в случае диадического анализа формальных понятий, мы будем пользоваться производными операторами (в диадическом случае это операторы Галуа) для удобства изложения и описания вычислений. Для краткости записи операторов обозначим триадический контекст несколько иначе (K_1,K_2,K_3,Y). Пусть {i,j,k}={1,2,3}, а j<k, тогда для $X \subseteq K_i$ и $Z \subseteq K_j \times K_k$, тогда (i)-производный оператор определяется следующими выражениями:

$$X \mapsto X^{(i)} = \{(a_j\ a_i) \in K_j \times K_k \mid (a_i, a_j, a_k) \in Y \text{ для всех } a_i \in X\}$$
$$Z \mapsto Z^{(i)} = \{a_j \in Ki \mid (a_i, a_j, a_k) \in Y \text{ для всех } (a_j, a_k) \in Z\}.$$

В случае бикластеризации мы предложили определение бикластера как пары виды (m′,g′) для всех пар $(g,m) \in I$. Плотным бикластером мы назвали пару (m′,g′), такую что плотность бикластера $\rho(m', g') = \dfrac{|m' \times g' \cap I|}{|m' \times g'|} \geq \rho_{\min}$, $\rho_{\min}$ — минимальное значение плотности, задаваемое пользователем.

В случае трикластеризации нам также понадобятся производные операторы, но только первого вида, сопоставляющие множеству элементов (например, объема) множества пар других элементов (первая компонента которых берётся из содержания, а вторая из модуса). Для упрощения нотации введем единое обозначение для такого оператора, действующего на трех различных множествах триконтекста. Мы будем применять этот оператор только к одноэлементным множествам, пусть $(g,m,b) \in I$, тогда

$$g' = \{(m, b) \in M \times B \mid (g, m, b) \in Y \text{ для всех } g \in G\}$$

[1] 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.

$m' = \{(g, b) \in G \times B \mid (g, m, b) \in Y$ для всех $m \in M\}$

$b' = \{(g, m) \in G \times M \mid (g, m, b) \in Y$ для всех $b \in B\}$.

Введём дополнительно бокс-операторы для элементов тройки $(g,m,b) \in I$:

$g^{\square} = \{g \mid (g, b) \in m'$ или $(g, m) \in b'\}$

$m^{\square} = \{m \mid (m, b) \in g'$ или $(g,m) \in b'\}$

$b^{\square} = \{b \mid (m, b) \in g'$ или $(g,b) \in m'\}$.

Трикластером назовем тройку вида $T=(g^{\square}, m^{\square}, b^{\square})$, где $(g,m,b) \in I$. Плотным трикластером назовем трикластер T, плотность которого превышает минимальное значение — порог, заданный пользователем, т.е. $\rho(T) = |g^{\square} \times m^{\square} \times b^{\square} \cap I| / |g^{\square} \times m^{\square} \times b^{\square}| \geq \rho_{min}$.

В таблице 1 приведён псевдокод предлагаемого нами алгоритма. Как и в двумерном случае, значение плотности трикластера лежит в интервале (0,1], а плотность трипонятия равна 1.

Таблица 1

Псевдокод алгоритма TRICL

TRICL((G,M,B,I) ρ_{min}**)**
Вход: формальный триконтекст (G,M,B,I) и значение минимальной допустимой плотности трикаластера.
Выход: TR – множество трикластеров.
For (g, m, b) **in** I $g' = \{(m, b) \in M \times B \mid (g, m, b) \in Y$ for all $g \in G\}$ $m' = \{(g, b) \in G \times B \mid (g, m, b) \in Y$ for all $m \in M\}$ $b' = \{(g, m) \in G \times M \mid (g, m, b) \in Y$ for all $b \in B\}$ **For** (g, m, b) **in** I $g^{\square} = \{g \mid (g, b) \in m'$ or $(g, m) \in b'\}$ $m^{\square} = \{m \mid (m, b) \in g'$ or $(g, m) \in b'\}$ $b^{\square} = \{b \mid (m, b) \in g'$ or $(g, b) \in m'\}$ if $\rho((g^{\square}, m^{\square}, b^{\square}) \geq \rho_{min}$ then $T=(g^{\square}, m^{\square}, b^{\square})$ Tr.Add(T) **Return** Tr

Утверждение 1. Для некоторого формального триконтекста количество порождаемых (плотных) трикластеров не превышает |I|.

Утверждение 2. Время порождения всех трикластеров не превышает O((|G||M|+|M||B|+|G||B|)|I|), а время порождения порождения всех плотных трикластеров O((|G||M|+|M||B|+|G||B|)|I||G||M||B|).

Что касается особенностей реализации алгоритма, то для избежания порождения дубликатов можно применять хеширование, т.е. для каждого нового трикластера находить значение хеш-функции, которое можно использовать в качестве ключа для хранения уникального трикластера в выходном массиве (словаре).

Гипотеза. Все формальные трипонятия исходного триконтекста (G, M, B, Y) содержаться в смысле покомпонентного вложения в некотором трикластере этого контекста.

Конечно, при увеличении минимального порога плотности до некоторого положительного значения данная гипотеза оказывается неверна. Однако формулировка гипотезы как положительного утверждения в настоящее время несколько поспешна и требует дополнительной проверки. У трикластера довольно простая геометрическая интерпретация — это параллелепипед в данных, в котором содержится «плотный» трехмерный крест, образованный пересечением трех меньших параллелепипедов с плотностью 1 в ячейке (g,m,b).

Обсудим другие полезные свойства трикластеров. Рассмотрим триконтекст (Пользователи, Тэги, Книги) в таблице 2. Трирешетка понятий такого контекста содержит $3^3=27$ различных понятий, в то время как TRICL находит только один трикластер ({Антон, Алёна, Настя},

{классика, судьба, чтиво}, {Золя.Западня, Достоевский.Крокодил, Островский.Гроза}) с плотностью ρ=0,89.

Таблица 2

«Слои» триконтекста (Пользователи, Тэги, Книги)

	классика	судьба	чтиво
Антон		X	X
Алёна	X	X	X
Настя	X	X	X
	Золя. Западня		

	классика	судьба	чтиво
Антон	X	X	X
Алёна	X		X
Настя	X	X	X
	Достоевский. Крокодил		

	классика	судьба	чтиво
Антон	X	X	X
Алёна	X	X	X
Настя	X	X	
	Островский. Гроза		

В случае реальных данных вычисление всех трипонятий становится практически невозможным, да и вряд ли будет иметь для экспертов особый смысл анализировать вручную количество понятий, превышающее несколько тысяч. Поэтому такие плотные трикластеры при разумном выборе порога являются средством описания трехкомпонентных сообществ пользователей в данных фолксономий. Что касается условия отбора плотных трикластеров, то, к сожалению, оно не является ни монотонным, ни антимонотонным (в более плотном трикластере может содержаться менее плотный и наоборот).

Реальные данные и эксперименты

Для экспериментов мы выбрали реальные данные[1] сервиса Bibsonomy.org, содержащие 816197 записей вида (пользователь, тэг, ресурс), представленные в качестве конкурсного набора данных на международной конференции European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases в 2008 году рамках состязания Discovery Challenge. Фактически мы имеем триконтекст, которому соответствует параллелепипед размерами |U||T||R|=2337×67464×28920. В качестве экспериментального оборудования нами использовался компьютер с процессором Pentium-DualCore с частотой 2,5ГГц и с 2ГБ оперативной памяти. В качестве языка программирования мы выбрали высокоуровневый мультипарадигменный язык Python версии 2.7.1, идеально подходящий для быстрого прототипирования реализаций алгоритмов.

Первые эксперименты на 100 тысячах исходных записей, которые соответствуют параллелепипеду размером 59×5823×28920=9935668440, показали, что алгоритм Tricl породил 4462 трикластера приблизительно за 3 часа (с учетом подсчета плотностей). Время вычисления всех трикластеров составило 1276 с, т.е. примерно 21 минуту (без учета подсчёта плотности). Как видим, отдельную проблему представляет собой точное вычисление плотности, которое сейчас осуществляется полной проверкой на принадлежность исходному триконтексту (отношению Y) тройки (g,m,b) — элемента конкретного трикластера T. Число таких проверок сопоставимо с размером параллелепипеда, соответствующего контексту. Альтернативным способом вычисления является рандомизированный приблизительный подсчет плотности, основанный на случайном выборе только N-ой доли ячеек трикластера для проверки принадлежности триконтексту. В случае хорошей корреляции этих приближенных значений с плотностью, такую оценку плотности можно использовать для отбора трикластеров по порогу её величины.

Интересным для исследователя является выполнение степенных законов распределения (Power Law) на данных бибсономии. Характерные примеры таких распределений наблюдают-

[1] http://www.kde.cs.uni-kassel.de/ws/rsdc08/

ся для посетителей сайтов, телефонных звонков, книг бестселлеров, интенсивности войн и т.п. Мы исследовали отдельно гистограммы распределения посетителей по количеству внесенных в систему пар (тэг, документ), тэгов — по количеству пар (пользователь, документ), ресурсов — по количеству пар (пользователь, тэг).

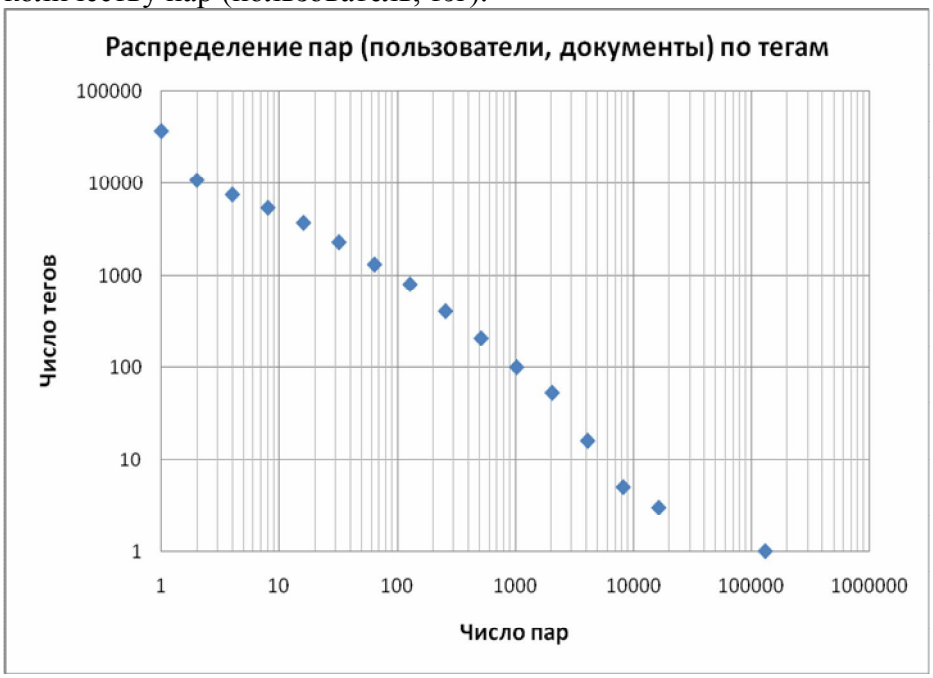

Рис. 1.

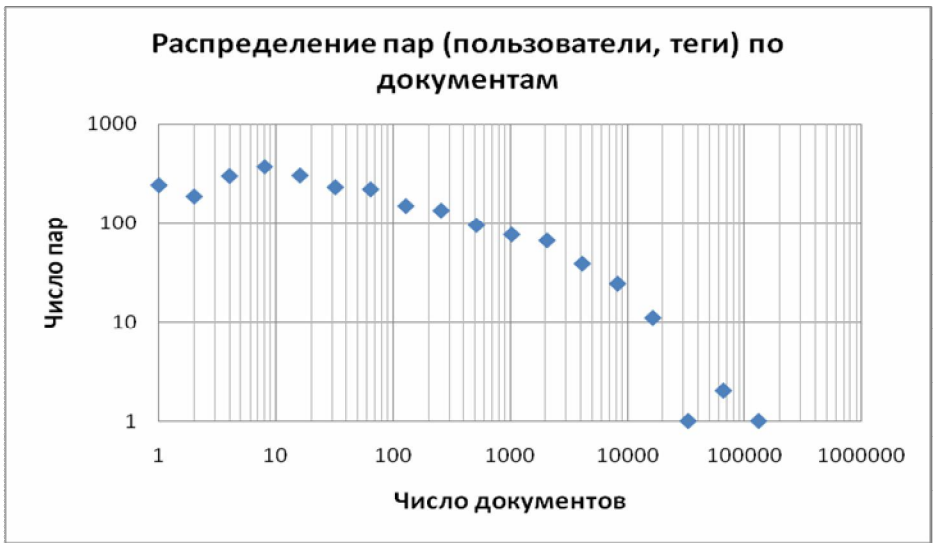

Рис. 2.

Для данных по количеству пар (пользователи, тэги) гипотеза о наличии степенного распределения не подтвердилась (рис. 2), хотя для распределения пар (пользователь, документ) этого нельзя с уверенностью сказать (рис. 1). Причиной этого могло послужить присутствие спамеров, которые генерировали высокочастотные тэги и документы. Распределение количества пар (пользователь, тэг), напротив демонстрирует степенной закон (явно выраженный прямолинейный участок в масштабе log-log, см рис. 3). Приблизительная оценка показателя *a* степени в законе распределения $p(x)=Cx^{-a}$ диаграммы на рис. 3 равна 1,5 (методика расчета изложена в работе Ньюмана[1]). Как видим, первые четыре наиболее высокочастотных тэга (табл. 3) выполняют служебную функцию и, скорее всего, порождаются системой автоматически.

[1] Newman M.E.J. Power laws, Pareto distributions and Zipf's law // Contemporary Physics. 2005. Vol. 46. No. 5.

Для данных по количеству пар (пользователи, тэги) гипотеза о наличии степенного распределения не подтвердилась (рис. 2), хотя для распределения пар (пользователь, документ) этого нельзя с уверенностью сказать (рис. 1). Причиной этого могло послужить присутствие спамеров, которые генерировали высокочастотные тэги и документы. Распределение количества пар (пользователь, тэг), напротив демонстрирует степенной закон (явно выраженный прямолинейный участок в масштабе log-log, см рис. 3). Приблизительная оценка показателя *a* степени в законе распределения $p(x)=Cx^{-a}$ диаграммы на рис. 3 равна 1,5 (методика расчета изложена в работе Ньюмана[1]).

Как видим, первые четыре наиболее высокочастотных тэга (табл. 3) выполняют служебную функцию и, скорее всего, порождаются системой автоматически.

Рис. 3.

Таблица 3

Top10 + 1 тэг

Тэг	Частота в парах присваиваний (пользователь, документ)
imported	66636
public	15666
system:imported	11294
nn	9147
video	7610
books	6214
software	5021
tools	4423
web2.0	4215
web	4071
blog	3439

Размеры трикластера — 2 пользователя, 27 тэгов и 8 документов, объём — 432, а плотность — 0.123. Даже такие, казалось бы, невысокие значения плотности, тем не менее, являются хорошим показателем релевантности трикластера, достаточно упомянуть, что плотность триконтекста, по которму был найден трикластер, составляет примерно 10^{-5}. Это позволяет нам с уверенностью говорит, что мы «нащупали сгусток плотности» в трикластере T. Интерпретация трикластера T довольно проста, мы видим все ресурсы (их идентификаторы) и тэги, которыми пользовались клиенты сервиса Bibsonomy с номерами 21 и 22 в некоторый временной промежуток, определяемый с момента запуска сервиса до появления 100000 записи.

[1] Newman M.E.J. Power laws, Pareto distributions and Zipf's law // Contemporary Physics. 2005. Vol. 46. No. 5.

Заключение

Результаты первых экспериментов позволят с уверенностью говорить, что метод находит релевантные трикластеры за приемлемое время (для сравнения, владельцы сайта сервиса Bibsonomy провели поиск крупных трипонятий на 600000 записях сервиса del.ici.ou.us примерно за 13 часов на мощном серверном оборудовании[1]). Однако данный метод нуждается в дальнейшем тестировании и доработке, как с точки зрения оценки мер качества кластеризации, так и с точки зрения оптимизации вычислений. Исследование характера распределений числа использования пользователем тэга или ресурса также является предметов дальнейшего исследования, т.к. следование степенным законам вполне оправдывает жадные стратегии отбора трикластеров по величине плотности. Метод может быть полезен специалистам в области прикладной математики и социологии, работающих с социальными Интернет-сервисами, в особенности, с такими, в которых могут возникать тернарные отношения на исследуемых объектах и ресурсах, а также при анализе сообществ, возникающих в фолксономиях. Например, существует термин «biclique communities», определяющий множество членов сообщества через множество их общих интересов, а бикластер, как известно, является ослаблением понятия биклики и может выполнять сходную функцию при выявлении сообществ. И в нашем случае мы имеем похожую ситуацию, в которой от вычислительно более сложно порождаемого объекта — трипонятия — мы переходим к вычислительно более эффективному, но менее жёсткому — трикластеру.

Язык синтаксических структур и структурирование текста

Кобзарева Татьяна Юрьевна, *РГГУ*

В статье рассматриваются некоторые проблемы синтаксического анализа русского предложения и специфика их решения в модульной системе автоматического анализа MARS[2], где поверхностно-синтаксическому моделированию структуры ситуаций предшествует сегментация предложения. В свете быстро совершенствующихся возможностей виртуального хранения информации важнейшим источником при сборе и структурировании социологической информации являются сообщения СМИ и новостных лент, другая информация из сети Internet, отчеты коммерческих и общественных организаций и т.п. Объем исходных документов исчисляется десятками тысяч, а при обработке коротких сообщений, новостей и статей в СМИ — сотнями тысяч источников. Все более актуальными становятся задачи автоматического поиска информации. Лингвистика, и в первую очередь компьютерная лингвистика, уже более 50-ти лет занимается компьютерным анализом текста. Результаты работ по автоматическому анализу, позволяющие учитывать при поиске информации связи слов в тексте, очень важны для решения задач сбора и структурирования информации, однако в настоящее время они практически не используются.

Информация линейной структуры предложения

При автоматическом анализе исходным объектом является последовательность слов и знаков препинания текста — его означающее. Фердинанд де Соссюр, один из самых блестящих лингвистов XX века, отмечал, что означающее, «являясь по своей природе воспринимаемым на слух, развертывается во времени и характеризуется заимствованными у времени признаками: а) оно обладает протяженностью, и б) эта протяженность имеет одно измерение — это линия. Об этом совершенно очевидном принципе сплошь и рядом не упоминают вовсе, по-видимому, именно потому, что считают его чересчур простым, между тем это весьма

[1] Jäschke R., Hotho A., Schmitz Ch., Ganter B., Stumme G. TRIAS - An Algorithm for Mining Iceberg Tri-Lattices // Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China. Hong Kong, 2006. P. 907–911.

[2] Система автоматического поверхностно-синтаксического анализа русского предложения MARS (modular analysis of Russian sentence), разрабатываемая автором в РГГУ при частичной поддержке РФФИ — грант № 09-06-00275-а.