Выявление нечетких кластеров ключевых понятий таксономии по коллекции текстов:

алгоритмы и экспериментальная верификация

аспирант 3 г.о. Фролов Д.С.
научный руководитель: Миркин Б.Г.

Структура доклада

1. Мотивация
2. Таксономии, определения и примеры
3. Аддитивная кластеризация алгоритмом FADDIS
4. Алгоритм обобщения кластеров на таксономии PAD
5. Эксперименты
6. Заключение

1. Мотивация: Поиск документов в коллекции

- **Скорость исполнения запроса +**
- **Нечеткость запроса +**
- **Интеллектуальный анализ результата с помощью отображения на таксономию ?**

2. Таксономии, определения и примеры

Таксономии - древообразные структуры классификации определенных наборов объектов

Пример - биологическая таксономия

Аддитивная нечеткая кластеризация (Mirkin, Nascimento, 2012)

Задано N-элементное множество T и матрица близости его элементов A = ($a_{tt'}$). Нечеткий кластер на T задается вектором степеней принадлежности U = (u_t), $t \in T$, $u_t \in [0, 1]$ и интенсивностью μ. Интенсивность масштабирует степени принадлежности для оценки их вклада в значения близостей.

Модель аддитивной кластеризации, K кластеров:

$$a_{tt'} = \sum_{k=1}^{K} \mu_k^2 u_{kt} u_{kt'} + e_{tt'}$$

u_k = (u_{kt}) - вектор степеней принадлежности для кластера k, μ - интенсивность

Аддитивная кластеризация: постановка задачи

$$a_{tt'} = \sum_{k=1}^{K} \mu_k^2 u_{kt} u_{kt'} + e_{tt'}$$

Задача поиска нечетких кластеров по матрице близости: задана $A = (a_{tt'})$, найти такие K кластеров (u_k), которые с интенсивностями μ_k минимизируют сумму квадратов ошибок $\sum_{t,t'} e_{tt'}^2$

Аддитивная кластеризация: решение

Кластеры ищутся пошагово, по одному.

Алгоритм FADDIS, основные принципы:

1. На каждом шаге для матрицы близости W = ($w_{tt'}$) решается задача поиска одного кластера методом наименьших квадратов. Минимизируется выражение

$$E = \sum_{t,t' \in T} (w_{tt'} - \xi u_t u_{t'})^2$$

$\xi > 0$, $u = (u_t)$

На начальном шаге W := A

2. После получения кластера u вычисляется новая матрица близости W как матрица остатков: W = $W - \mu^2 \mathbf{u}\mathbf{u}'$

Процесс повторяется с Шага 1 для новой матрицы W.

Выражение для ξ:

$$\xi = \frac{\mathbf{u}'W\mathbf{u}}{(\mathbf{u}'\mathbf{u})^2} \qquad \xi \geq 0$$

$$E = \sum_{t,t' \in T} w_{tt'}^2 - \xi^2 \sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2 = S(W) - \xi^2 (\mathbf{u}'\mathbf{u})^2$$

$$G(\mathbf{u}) = \xi^2 (\mathbf{u}'\mathbf{u})^2 = \left(\frac{\mathbf{u}'W\mathbf{u}}{\mathbf{u}'\mathbf{u}} \right)^2 \qquad S(W) = G(\mathbf{u}) + E$$

После преобразований:

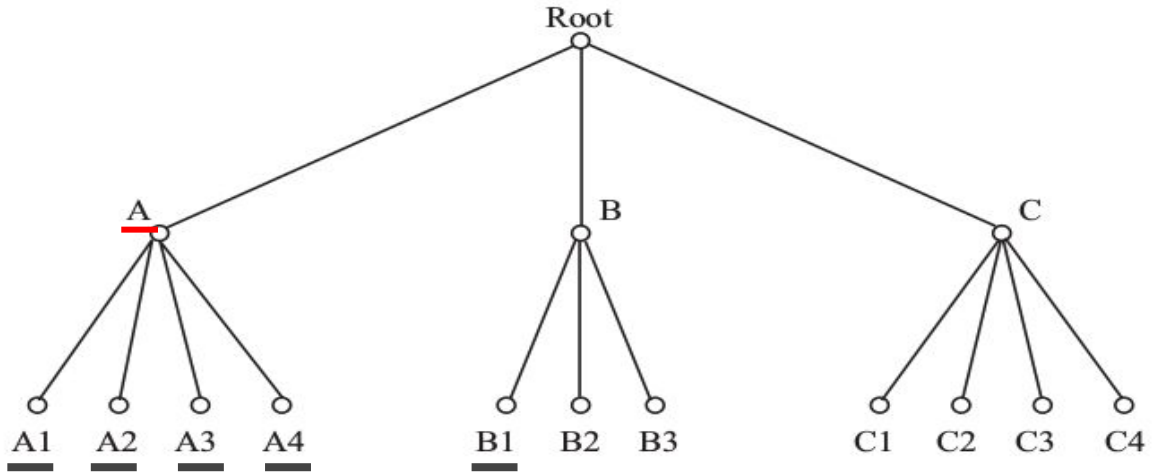
$$g(\mathbf{u}) = \xi \mathbf{u}'\mathbf{u} = \frac{\mathbf{u}'W\mathbf{u}}{\mathbf{u}'\mathbf{u}}$$

- отношение Рэлея, в данном случае его максимум достигается при выборе вектора u, соответствующего максимальному собственному числу матрицы W, если на u нет дополнительных ограничений. Именно так задача и решается, а потом для полученного вектора z выполняется преобразование координат: $u_t = \max(0, z_t)$

Условие остановки алгоритма возможно реализовать по условиям:

1. $\xi < 0$
2. Вклад очередного полученного кластера в матрицу близости слишком незначительный (эта граница, например, может быть определена заранее)
3. Норма матрицы остатков стала ниже предопределенного значения
4. Получено заранее определенное число кластеров

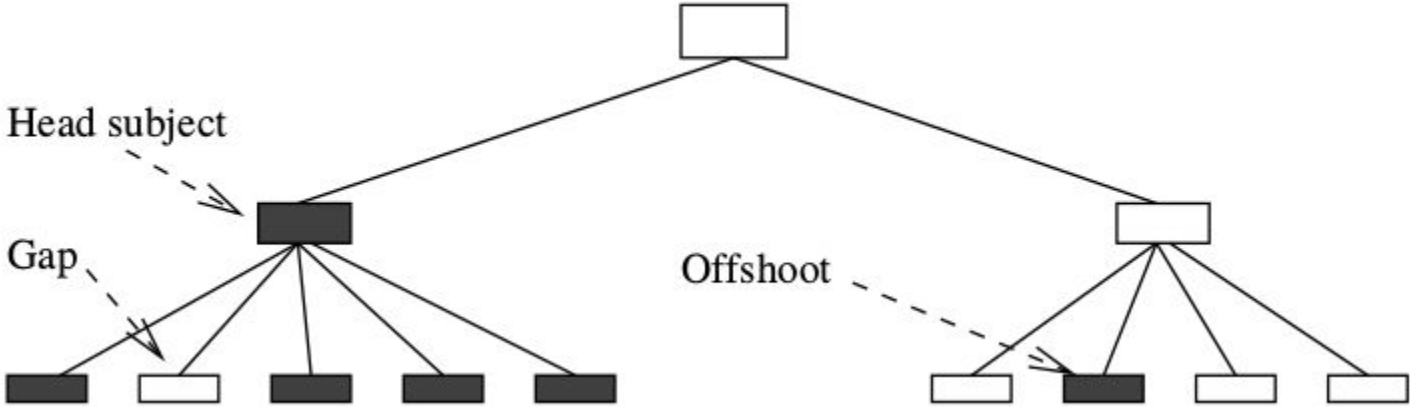
Интерпретация кластеров с помощью подъема в более общие вершины (Mirkin, Fenner, Nascimento, in progress)



Если известна таксономия и получен кластер, в интерпретации удобно заменить листья на узлы более высокого уровня, если это возможно:

(A1, A2, A3, A4, B1) => (A), B1 - выброс

Логика подхода



Возможно наличие элементов типа gap - пропуск, или offshoot - выброс.

Обобщающая вершина - головной таксон (head subject).

Определения

Задано дерево таксономии T, аннотированное ключевыми словами (в узлах содержатся названия таксонов). Множество листьев дерева - I, внутренних узлов T - I. Для каждого узла t обозначим x(t) - множество его прямых потомков. Нечеткий кластер на I: $S_u = \{i \in I: u(i) > 0\}$, где u(i) - значение функции принадлежности к кластеру.

S_u называется поддержкой кластера u.

T(t) - поддерево с корнем в узле t. Множество листьев поддерева называется листовым кластером узла t - I(t).

Если задан нечеткий кластер u на листьях дерева T, можно определить головные таксоны в кластере u (возможно, с погрешностями).

Головными таксонами будут такие внутренние вершины h, для которых листовой кластер совпадает (в пределах допустимой погрешности) с поддержкой S_u

Два типа ошибок: gap и offshoot

Узел $t \in T$ называется u-иррелевантным, если его листовой кластер $l(t)$ не пересекается с поддержкой S_u. Если узел u-иррелевантен, очевидно, что все его потомки тоже u-иррелевантны. Это означает, что если головной таксон h потерян в узле t, то он отсутствует и во всех его потомках.

h-gap - такой узел g в $T(h)$ ($g \neq h$), в котором такая потеря произошла, что равнозначно условию g - максимальный u-иррелевантный узел в том смысле, что его родитель не u-иррелевантен. Наоборот, назначение головного таксона h в узле t будет являться приобретением h.

$G(h)$ - множество всех h-gap.

h-offshoot - листовой узел $i \in S_u$, не покрытый h, то есть $i \notin I(h)$

Также вводится “значимость” gap - v(g): например, v(g) = u(par(g)), где - par(g) - родительский узел.

Алгоритм PAD (Mirkin, Fenner, Nascimento)

Parsimonious Approximate Descriptor

Алгоритм строит множество головных таксонов H, **минимизирующее** штрафную функцию:

$$p(H) = \sum_{h \in H-I} u(h) + \sum_{h \in H-I} \sum_{g \in G(h)} \lambda v(g) + \sum_{h \in H \cap I} \gamma u(h)$$

λ - штраф за gap

γ - штраф за offshoot, 1 - штраф за головную вершину

I - множество листьев дерева

Предварительные шаги алгоритма:

1. Аннотировать все узлы дерева таксономии значениями функции принадлежности к кластеру. Поскольку листовые узлы уже аннотированы, действие выполняется для внутренних узлов.
2. Удалить из дерева таксономии все не максимальные u-иррелевантные узлы, то есть потомки гэпов.
3. Вычислить множества гэпов G(t) во всех узлах t дерева T.
4. Для всех узлов вычислить суммы значимостей их гэпов:

$$V(t) = \sum_{g \in G(t)} v(g)$$

Алгоритм вычисляет для каждого узла дерева таксономии два множества:

1. H(t) - множество приобретенных в узле головных таксонов
2. L(t) - множество потерянных таксонов

Штраф за операции будет храниться в p(t).

Алгоритм рекурсивно вычисляет все эти значения, начиная с листьев дерева и заканчивая корнем.

Для каждого узла дерева возможны два случая: головной таксон либо приобретается (1), либо нет (2).

Случай (1) - головной таксон приобретается. В этом случае сохранять значения списков H и L всех родительских узлов не требуется. Новые значения определяются как:

$$H(t) = \{t\}$$

$$L(t) = G(t)$$

$$p(t) = u(t) + \lambda V(t)$$

Случай (2) - головной таксон не приобретается. В этом случае списки H и L определяются как объединения значений списков всех дочерних узлов:

$$H(t) = \bigcup_{w \in \chi(t)} H(w)$$

$$L(t) = \bigcup_{w \in \chi(t)} L(w)$$

$$p(t) = \sum_{w \in \chi(t)} p(w)$$

Выбирается случай 1 или 2, в зависимости от штрафа p(t) - где меньше.

Алгоритм PAD - псевдокод

- **INPUT:** u, T
- **OUTPUT:** $H = H(root),\ L = L(root),\ p = p(root)$

I Base Case

for each leaf $i \in I$

if $u(i) > 0$

$H(i) = \{i\}$

$L(i) = \oslash$

$p(i) = \gamma u(i)$

else

$H(i) = \oslash$

$L(i) = \oslash$

$p(i) = 0$

II Recursion

if $u(t) + \lambda V(t) \leq \sum_{w \in \chi(t)} p(w)$

$H(t) = \{t\}$

$L(t) = G(t)$

$p(t) = u(t) + \lambda V(t)$

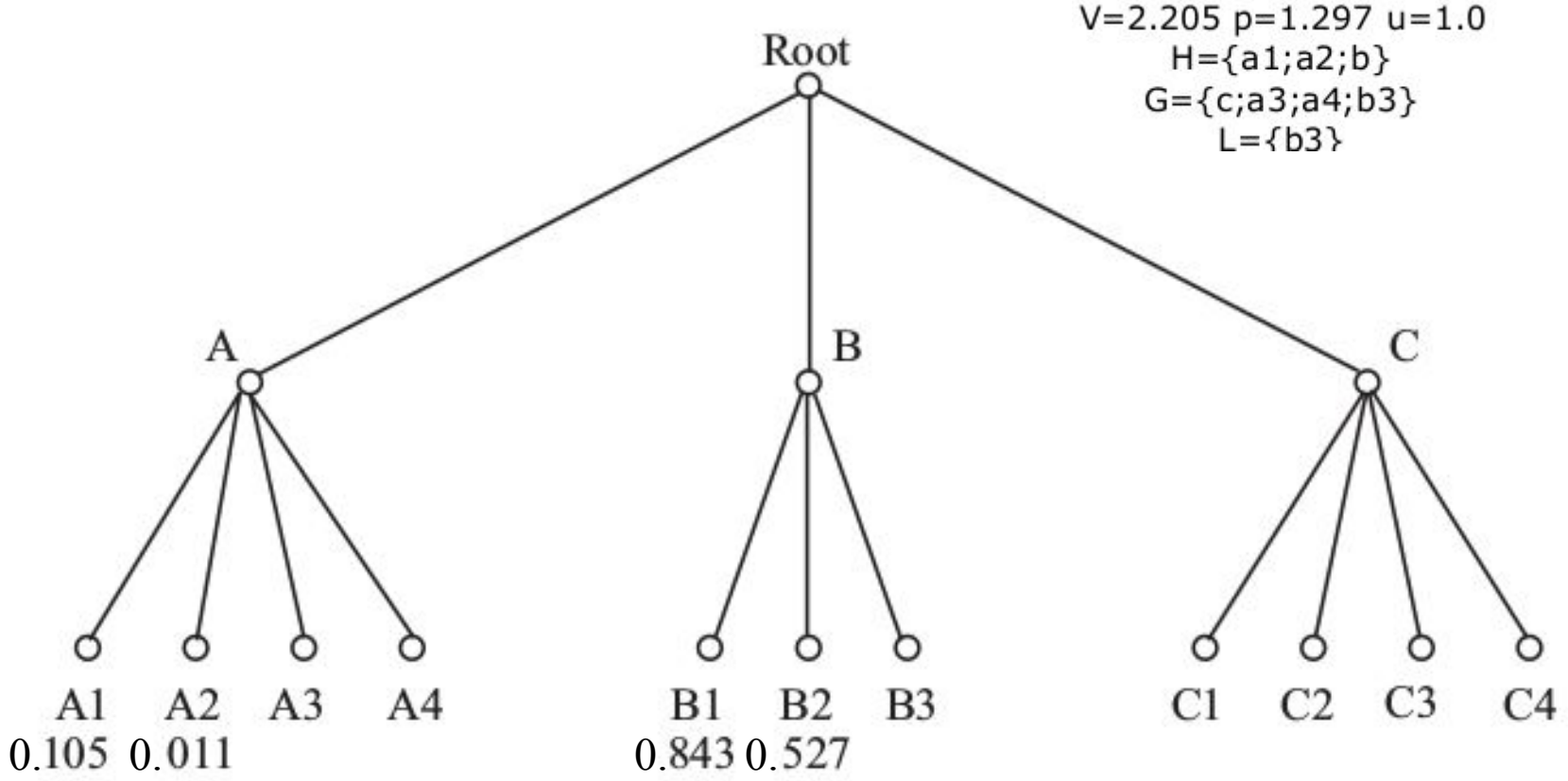
else

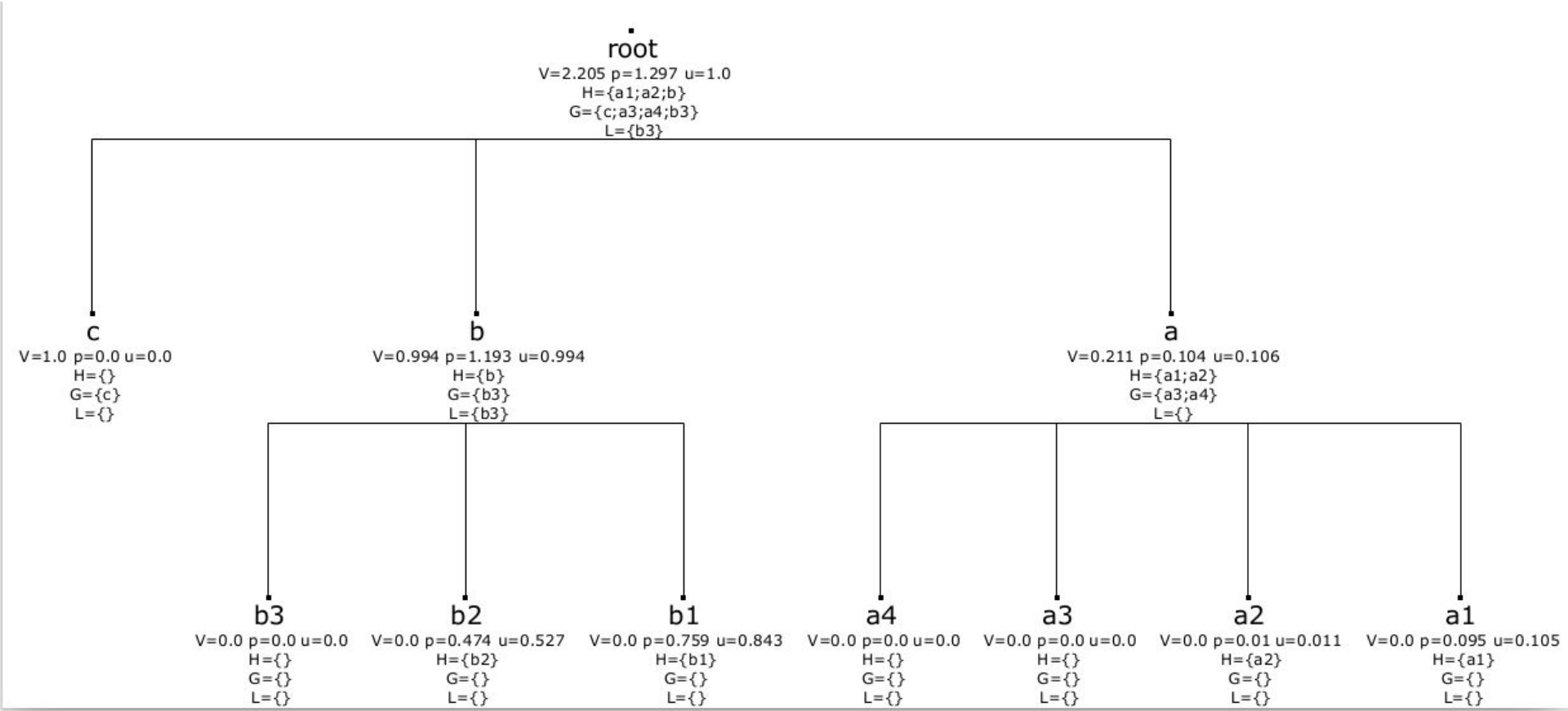
$H(t) = \bigcup_{w \in \chi(t)} H(w)$

$L(t) = \bigcup_{w \in \chi(t)} L(w)$

$p(t) = \sum_{w \in \chi(t)} p(w)$

PAD - примеры использования




root
V=2.205 p=1.297 u=1.0
H={a1;a2;b}
G={c;a3;a4;b3}
L={b3}
c
V=1.0 p=0.0 u=0.0
H={}
G={c}
L={}
b
V=0.994 p=1.193 u=0.994
H={b}
G={b3}
L={b3}
a
V=0.211 p=0.104 u=0.106
H={a1;a2}
G={a3;a4}
L={}
b3
V=0.0 p=0.0 u=0.0
H={}
G={}
L={}
b2
V=0.0 p=0.474 u=0.527
H={b2}
G={}
L={}
b1
V=0.0 p=0.759 u=0.843
H={b1}
G={}
L={}
a4
V=0.0 p=0.0 u=0.0
H={}
G={}
L={}
a3
V=0.0 p=0.0 u=0.0
H={}
G={}
L={}
a2
V=0.0 p=0.01 u=0.011
H={a2}
G={}
L={}
a1
V=0.0 p=0.095 u=0.105
H={a1}
G={}
L={}

λ=0.2, γ = 0.9

5. Эксперименты

Были проведены эксперименты на коллекции статей с сайта SpringerOpen (springeropen.com) - коллекция содержит журнальные публикации по разным тематикам

1. Загружено 20 тыс. статей журналов из раздела ComputerScience.
2. Сделана выборка статей, содержащих в аннотации хотя бы одно из слов списка [“cluster”, “data analysis”, “regression”, …] - получено 357 статей

3. Взят фрагмент таксономии ACM (ACM Computing Classification System). Листовых таксонов в нем - 317.

1.Theory of computation

1.1. Theory and algorithms for application domains

1.1.1. Machine learning theory

1.1.1.1. Sample complexity and generalization bounds

1.1.1.2. Boolean function learning

1.1.1.3.Unsupervised learning and clustering

1.1.1.4. Kernel methods

1.1.1.4.1. Support vector machines

1.1.1.4.2. Gaussian processes

...

4. С помощью метода аннотированного суффиксного дерева определения релевантности строки тексту построена матрица $A = (A_{ij})$ размера 317*357 - матрица, содержащая значения релевантностей тематических единиц таксономии статьям.

5. Получены кластеры тематических единиц следующими методами:

1. Псевдо-обратное преобразование Лапласа + FADDIS
2. Спектральный метод
3. Consensus k-Means

6. К кластерам применен алгоритм PAD. Пример - результат поиска головных таксонов для кластера, содержащего 10 тем:

massive data clustering 0.3015
consensus clustering 0.3015
fuzzy clustering 0.3015
additive clustering 0.3015
feature weight clustering 0.3015
conceptual clustering 0.3015
biclustering 0.3015
graph based conceptual clustering 0.3015
trajectory clustering 0.3015
spectral clustering 0.3015

Множество головных таксонов **H = {clustering, graph based conseptual clustering, trajectory clustering, clustering and classification, spectral methods}**

Полное дерево: http://dmitryfrolov.com/pdf/tree_1.pdf

6. Заключение. Применение в рекламном таргетинге?

Алгоритмическая модель закупки рекламы (programmatic, RTB) предполагает закупку рекламодателем сегментов пользователей с определенными свойствами. Примеры запросов: “Владельцы автомобилей с дизельными двигателями”, “Люди младше 20 лет”. Сегменты отображаются на таксономию, например, какую-нибудь из зарубежных стандартов IAB Taxonomies (как для сегментов пользователей, так и для контента сайтов)

Алгоритм получения головных таксонов способен улучшить точность рекламного таргетинга - требуется изучение применимости.

Литература

1. Nascimento S., Mirkin B., Moura-Pires F. A fuzzy clustering model of data and fuzzy c-means //Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on. – IEEE, 2000. – Т. 1. – С. 302-307.
2. Mirkin B., Nascimento S. Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices //Information Sciences. – 2012. – Т. 183. – №. 1. – С. 16-34.
3. Mirkin B. G. et al. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes //BMC evolutionary biology. – 2003. – Т. 3. – №. 1. – С. 2.
4. Mirkin B. et al. Constructing and mapping fuzzy thematic clusters to higher ranks in a taxonomy //International Conference on Knowledge Science, Engineering and Management. – Springer Berlin Heidelberg, 2010. – С. 329-340.
5. Миркин Б. Г., Черняк Е. Л., Чугунова О. Н. Метод аннотированного суффиксного дерева для оценки степени вхождения строк в текстовые документы //Бизнес-информатика. – 2012. – №. 3 (21).

Спасибо за внимание!