

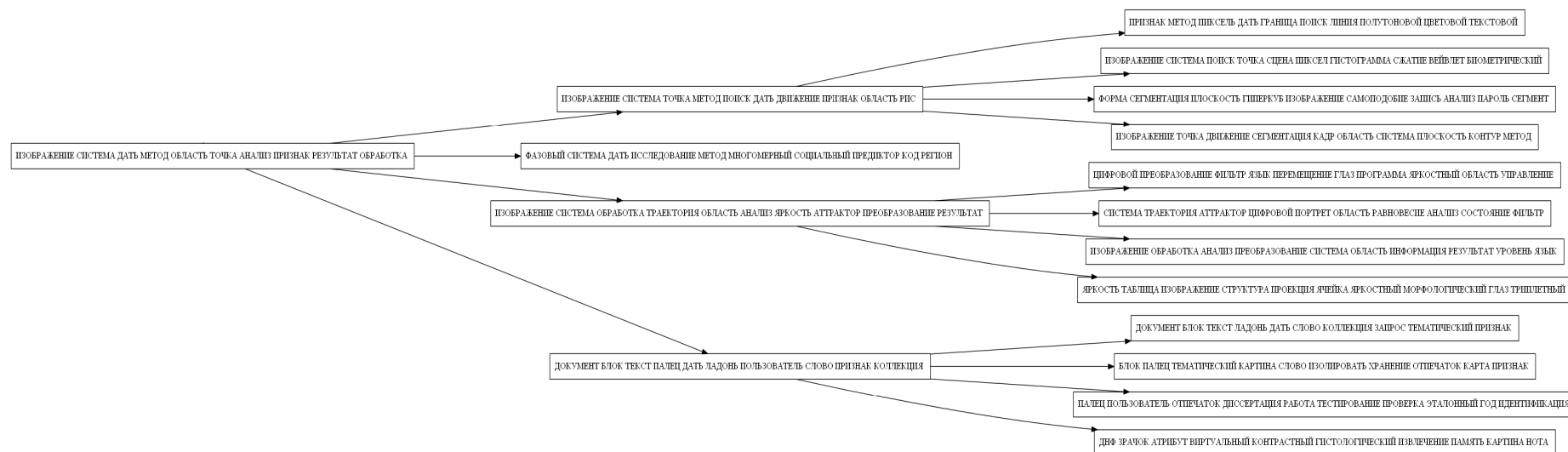
Иерархические Тематические Модели

Надежда Чиркова

Научный семинар «Машинное обучение и информационный поиск»
ШАД Яндекс, 30 сентября 2014

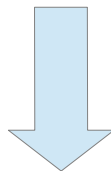
Иерархическая тематическая модель — это дерево тем.

Каждая тема — это набор **слов**, **авторов**, **документов**
+ множество **подтем**



Фрагмент иерархии, посвященный теме «Изображения»

Человеку гораздо **легче интерпретировать иерархию** тем, чем
плоскую тематическую модель

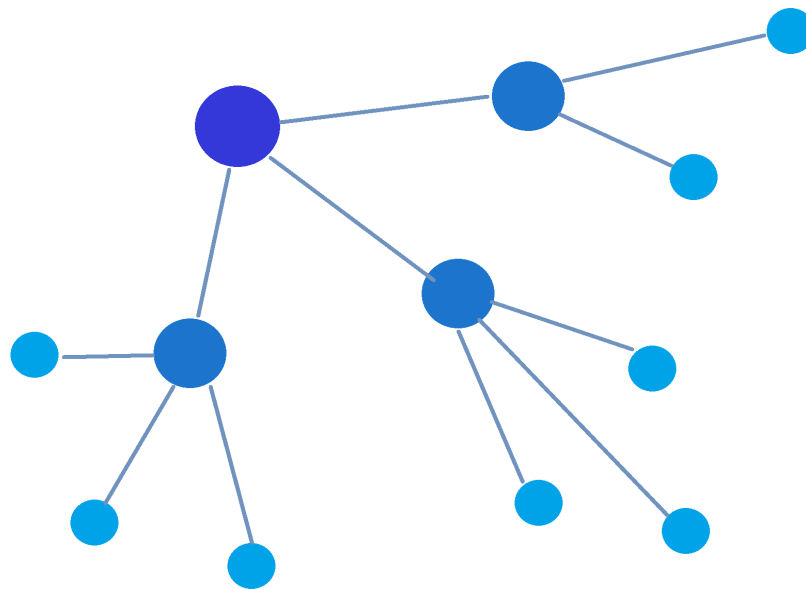


Можно сделать удобный и понятный

Навигатор по коллекции документов:

topnav.esy.es

Навигатор по коллекции статей
Конференций
«Интеллектуализация обработки данных»
И «Математические методы распознавания образов»



Тема 0

Подтема 0

задача	▲	Дюкова Е. В.	▲	Распознавание	▲
алгоритм	■	Кельманов А. В.	■	промоторов ДНК	■
решение	■	Воронцов К. В.	■	на основе	■
множество	▼	Стрижов В. В.	▼	электростатическо	▼
число		Дорофеюк Ю. А.			

Подтема 1

объект	▲	Моттль В. В.	▲	Байесовский	▲
алгоритм	■	Красоткина О. В.	■	подход к задаче	■
признак	■	В. Воронцов К.	■	обучения	■
функция	▼	В. Дюкова Е. В.	▼	распознаванию	▼
оценка		Середин О. С.			

Подтема 2

изображение	▲	Неймарк Ю. И.	▲	Методы анализа	▲
точка	■	Леухин А. Н.	■	структурной	■
сигнал	■	Котельников И. В.	■	плазменной	■
система	▼	Немирко А. П.	▼	турбулентности	▼
параметр		П. Кий К. И.			

Фоновая тема 3

сигнал	▲	Панкратов А. Н.	▲	Внедрение	▲
документ	■	Чалей М. Б.	■	системы	■
коэффициент	■	Назипова Н. Н.	■	<<Антиплагиат>>	■
текст	▼	Тетуев Р. К.	▼	в Российской	▼
разложение		Дедус Ф. Ф.			

Фоновая тема 4

задача	▲	Кельманов А. В.	▲	Разработка	▲
объект	■	Моттль В. В.	■	средств анализа	■
дать	■	Середин О. С.	■	астрономических	■
изображение	▼	Дорофеюк А. А.	▼	баз данных	▼
алгоритм		Сулимова В. В.		методами	

Слова(11645)

задача
алгоритм
объект
дать
метод
изображение
значение
функция
множество
являться
решение
модель
работа
оценка
признак
число
точка
система
класс
результат
вектор
параметр
случай
распознавание
анализ
образ
мочь
следовать
получить
выборка
матрица
вид
качество
соответствовать
иметь
использовать
последовательность
набор
пространство
элемент
построение
подход
время

Авторы(978)

Моттль В. В.
Кельманов А. В.
Воронцов К. В.
Красоткина О. В.
Дорофеюк А. А.
Дюкова Е. В.
Середин О. С.
Дорофеюк Ю. А.
Панкратов А. Н.
Сулимова В. В.
Пытьев Ю. П.
Дедус Ф. Ф.
Тетуев Р. К.
Двоенко С. Д.
Ветров Д. П.
Ивахненко А. А.
Федотов Н. Г.
Татарчук А. И.
Неймарк Ю. И.
Мясников В. В.
Немирко А. П.
Стрижов В. В.
Кропотов Д. А.
Леухин А. Н.
Янковская А. Е.
Ланге М. М.
Теклина Л. Г.
Пятков М. И.
Инякин А. С.
Назипова Н. Н.
Сенько О. В.
Копылов А. В.
Чуличков А. И.
Котельников И. В.
Махортых С. А.
Дьяконов А. Г.
Хачай М. Ю.
Устинин М. Н.
Чалей М. Б.
Ушмаев О. С.
Козодеров В. В.
Ботов П. В.
Кузнецов М. И.

Статьи(983)

- Разработка средств анализа астрономических баз данных методами когнитивной визуализации
- Внедрение системы <<Антиплагиат>> в Российской государственной библиотеке
- Распознавание промоторов ДНК на основе электростатического потенциала
- Применение методов распознавания образов в системе управления коллекциями графических документов
- Формализация задачи распознавания последовательности состояний сложного источника
- Прикладные технологии распознавания количественных характеристик растительности
- Построение параметрического портрета динамической системы на основе синдромальных представлений
- Автоматизированная *
- СРС-методы как методы интеллектуального анализа данных при исследовании реальных хаотических процессов
- Байесовский подход к задаче обучения распознаванию образов в нестационарной генеральной совокупности
- Анализ техники живописи по изображениям в задачах атрибуции. Обзор
- Методы анализа структурной плазменной турбулентности
- О моделировании мышления в реальном мире
- Параметрическое семейство гранично-скелетных моделей формы
- Новая технология численного исследования динамических систем

Тема 0/2

Надтема

Подтема 0

изображение	▲ Кий К. И. Немирко	Виртуальные
точка	▲ А. П. Пытьев Ю. П.	границные
сигнал	▲ Калиниченко А. Н.	кривые; подход к
метод	▲ Манило Л. А.	анализу движения
область		

Подтема 1

параметр	▲ Котельников И.	▲ Исследование
система	▲ В. Неймарк Ю.	математической
область	▲ И. Теклина Л. Г.	модели
состояние	▲ Леухин А. Н.	экологической
распознавание	▲ Кий К. И.	

Подтема 2

сигнал	▲ Леухин А. Н.	▲ Методы анализа
последовательн	▲ Мясников В. В.	структурной
система	▲ Немирко А. П.	плазменной
цифровой	▲ Васин Ю. Г.	турбулентности
	▲ Лебедев Л. И.	

Подтема 3

изображение	▲ Ташлинский А.	▲ Идентификация
точка	▲ Г. Бакина И. Г.	личности по
параметр	▲ Жарких А. А.	форме ладони и
признак	▲ Мясников В. В.	голосу Оценка
оценка	▲ Федотов Н. Г.	качества

Фоновая тема 4

опасный	▲ Леухин А. Н.	▲ Исследование
параметризация	▲ Дорофеук А. А.	эффективности
авто	▲ Неймарк Ю. И.	регрессионной
штраф	▲ Чуличиков А. И.	модели
составить	▲ Теклина Л. Г.	

Слова(3551)

изображение
точка
сигнал
система
параметр
область
распознавание
последовательность
рис
состояние
характеристика
контур
уровень
исследование
преобразование
обработка
движение
этап
граница
траектория
кривая
фрагмент
длина
изменение
определение
размер
фазовый
ряд
локальный
яркость
координата
цифровой
статистический
спектр
закономерность
фильтр
шум
момент
цветовой
интервал
выделение
определяется

Авторы(676)

Неймарк Ю. И.
Леухин А. Н.
Котельников И. В.
Немирко А. П.
Кий К. И.
Теклина Л. Г.
Мясников В. В.
Дорофеук А. А.
Федотов Н. Г.
Бакина И. Г.
Ивановский С. А.
Дорофеук Ю. А.
Чуличиков А. И.
Марьяскин Е. Л.
Фурман Я. А.
Ташлинский А. Г.
Манило Л. А.
Романов С. В.
Пытьев Ю. П.
Лебедев Л. И.
Котов Ю. Б.
Козодеров В. В.
Ветров Д. П.
Обухов Ю. В.
Демин Д. С.
Жарких А. А.
Тюкаев А. Ю.
Дмитриев Е. В.
Васин Ю. Г.
Анциперов В. Е.
Парсаев Н. В.
Махортых С. А.
Кропотов Д. А.
Мокшанина Д. А.
Шпехт И. А.
Калиниченко А. Н.
Волкова С. С.
Панкратов А. Н.
Сенько О. В.
Кондранин Т. В.
Хашин С. И.
Рогов А. И.

Статьи(550)

- Методы анализа структурной плазменной турбулентности
- Построение параметрического портрета динамической системы на основе синдромальных представлений
- Исследование математической модели экологической системы на основе синдромальных представлений распознавания образов
- Формализация задачи распознавания последовательности состояний сложного источника
- Новая технология численного исследования динамических систем методами распознавания образов
- Виртуальные граничные кривые: подход к анализу движения
- Модифицированный метод геометризованных гистограмм и его применение
- Метод *
- Оценка качества JPEG2000 изображений
- Идентификация личности по форме ладони и голосу
- Методика интеллектуального анализа квазипериодических биосигналов
- Метод сравнения формы ладоней при наличии артефактов
- Разработка математических методов формализации профессионального знания врача
- Метод распознавания размытых штрихкодов на мобильных устройствах без автофокусировки
- Идентификация модели ладони по серии её снимков в разных положениях
- Выделение радужки методом

Тема 0/2/3

Надтема

Подтема 0

точка	▲ Ташлинский А. Г.	▲ Оценка качества
параметр	Г. Мясников В.	JPEG2000
оценка	В. Лепский А. Е.	изображений
кривизна	Хрящев В. В.	Оптимальный
локальный	Рейер И. А.	▼

Подтема 1

изображение	▲ Ташлинский А. Г.	▲ Метод сравнения
признак	Г. Федотов Н. Г.	формы ладоней
отсчет	Романов С. В.	при наличии
ладонь	Левашкина А.	артефактов
палец	О. Поршнева С.	▼

Подтема 2

точка	▲ Жарких А. А.	▲ Идентификация
распознавание	Бакина И. Г.	личности по
метод	Лепский А. Е.	форме ладони и
направление	Леухин А. Н.	голосу
зрочок	Потехин Е. Н.	Выделение

Фоновая тема 3

оцениваться	▲ Ташлинский А. Г.	▲ Непрерывный
задача	Г. Федотов Н. Г.	метод
множество	Леухин А. Н.	вычисления
решение	Жарких А. А.	морфологического
класс	Романов С. В.	▼

Слова(581)

изображение
точка
параметр
признак
оценка
ладонь
распознавание
координата
фрагмент
преобразование
локальный
рис
кривая
угол
кривизна
результат
контур
отсчет
граница
пиксель
яркость
направление
размер
мочь
обнаружение
лицо
алгоритм
спектр
палец
слово
поверхность
функция
человек
зрочок
характеристика
цвет
длина
модель
перенос
петроглиф
качество
блок

Авторы(264)

▲ Ташлинский А. Г.
Бакина И. Г.
Жарких А. А.
Мясников В. В.
Федотов Н. Г.
Лепский А. Е.
Леухин А. Н.
Хрящев В. В.
Левашкина А. О.
Поршнева С. В.
Романов С. В.
Мокшанина Д. А.
Фурман Я. А.
Быстров М. Ю.
Мельниченко А. С.
Потехин Е. Н.
Харитонов А. В.
Кириков П. В.
Рогов А. А.
Мурашов Д. М.
Макарова Е. Ю.
Чочиа П. А.
Рогова К. А.
Рейер И. А.
Жукова К. В.
Ветров Д. П.
Чуликов А. И.
Дмитриев Е. В.
Козодеров В. В.
Жизняков А. Л.
Середин О. С.
Каримов М. Г.
Тухтасинов М. Т.
Гетманов В. Г.
Роженцов А. А.
Наумов А. С.
Де Ванса Викраматне В. К.
Гальяно Ф. Р.
Дышкант Н. Ф.
Егоров В. Д.
Нагапетян В. Э.
Ипатов Ю. А.
Козловский А. В.

Статьи(172)

- Идентификация личности по форме ладони и голосу
- Оценка качества JPEG2000 изображений
- Метод сравнения формы ладоней при наличии артефактов
- Выделение радужки методом оптимизации кругового пути
- Подход к измерению активности выброса по данным мониторинга радиационной обстановки
- Методика привязки изображений в условиях интенсивных помех
- Оптимальный выбор параметров в упрощенной схеме детектора Харриса
- Вероятности распознавания направления переноса в одной модели случайного движения точки на плоскости
- Оценка кривизны методом усреднения локально-интерполяционных оценок
- Моделирование видеoinформационного тракта оптико-электронных систем дистанционного зондирования
- Выявление <<следов>> применения алгоритмов цифровой обработки на изображениях
- Обнаружение информативных фрагментов в задаче оценки качества изображений
- Трейс-преобразование как источник признаков распознавания
- Обнаружение нехарактерных участков на изображении с помощью фрактальных признаков самоподобия
- Распознавание *
- Обнаружение точек на контурах теней объекта, сопряженных с точками на его поверхности
- Анализ работы алгоритмов выделения признаков изображений

Тема 0/0

Надтема

Подтема 0

признак	▲ Чалей М. Б.	▲ Скрытая
последовательность	Кутыркин В. А.	периодичность в
строка	Назипова Н. Н.	кодирующих
лнк	Янковская А. Е.	последовательностей
	Тетев Р. К.	

Подтема 1

множество	▲ Торшин И. Ю.	▲ Кооперативные
задача	Папилин С. С.	стратегии для
условие	Дорофеев Н. Ю.	возможностей
игрок	Дьяконов А. Г.	моделей
стратегия	Иофина Г. В.	

Подтема 2

матрица	▲ Дюкова Е. В.	▲ * Об
алгоритм	Нефёдов В. Ю.	асимптотически
признак	Инякин А. С.	эффективном
задача	Сотнезов Р. М.	поиске
столбец	Иофина Г. В.	

Подтема 3

модель	▲ Кумсков М. И.	▲ Построение и
выборка	Моттль В. В.	использование
слово	Мучник И. Б.	адаптивных
функция	Сулимова В. В.	распознающих
класс	Пытьев Ю. П.	

Подтема 4

покрытие	▲ Дюкова Е. В.	▲ О некоторых
число	Генрихов И. Е.	аспектах
класс	Инякин А. С.	интеллектуального
ряд	Емельянов Г. М.	анализа пучков
понятие	Майсурадзе А.	

Слова(3181)

алгоритм
матрица
множество
модель
признак
построение
последовательность
строка
набор
условие
описание
ряд
шаг
столбец
длина
граф
слово
покрытие
понятие
построить
структура
возможность
определить
пар
переменный
временный
функционал
называться
содержать
закономерность
оператор
случайный
операция
максимальный
днк
стратегия
алгебраический
скрыть
сеть
нечеткий
конечный
булев
называть

Авторы(611)

Дюкова Е. В.
Моттль В. В.
Чалей М. Б.
Торшин И. Ю.
Кутыркин В. А.
Инякин А. С.
Нефёдов В. Ю.
Емельянов Г. М.
Стризов В. В.
Сотнезов Р. М.
Дьяконов А. Г.
Генрихов И. Е.
Сулимова В. В.
Михайлов Д. В.
Дорофеев Ю. А.
Дорофеев А. А.
Пытьев Ю. П.
Янковская А. Е.
Дорофеев Н. Ю.
Майсурадзе А. И.
Назипова Н. Н.
Кумсков М. И.
Красоткина О. В.
Воронцов К. В.
Федотов Н. Г.
Иофина Г. В.
Тетев Р. К.
Дедус Ф. Ф.
Разин Н. А.
Максимов Ю. В.
Мучник И. Б.
Ивахненко А. А.
Колесниченко А. С.
Хачай М. Ю.
Панкратов А. Н.
Гуров С. И.
Папилин С. С.
Филиппенков Н. В.
Чехович Ю. В.
Покровская И. В.
Пятков М. И.
Прохоров Е. И.
Середин О. С.

Статьи(551)

- Построение и использование адаптивных распознающих моделей
- Об асимптотически оптимальном построении элементарных классификаторов
- Кооперативные стратегии для возможных моделей биматричных игр
- О корректном понижении значности данных в задачах распознавания
- Потенциальные функции на множестве аминокислот на основе модели эволюции М. Дэйхофф
- Критерии локальной разрешимости и регулярности для исследования аминокислотных последовательностей
- Скрытая периодичность в кодирующих последовательностях ДНК
- О свойствах задач и алгоритмов разметки элементов точечных конфигураций
- Построение и исследование полиномиальных алгоритмов для логического анализа данных
- О сложности логического анализа данных в распознавании
- Об асимптотически эффективном поиске конъюнктивных закономерностей
- О некоторых аспектах интеллектуального анализа пучков временных рядов
- Скрытая профильная периодичность как новый тип периодичности генома
- Пространство формализации изображений
- *
- Семантическая схожесть текстов в задаче автоматизированного контроля знаний
- Сравнение эвристических алгоритмов

Тема 0/0/0

Надтема

Подтема 0

строка	▲ Чалей М. Б.	▲ Скрытая
последовательность	Кутыркин В. А.	периодичность в
скрыть	Назипова Н. Н.	кодированных
лнк	Тетуев Р. К.	последовательность
	Федотов Н. Г.	

Подтема 1

матрица	▲ Дедус Ф. Ф.	▲ Оптимизация *
последовательность	Панкратов А. Н.	Интеллектуальный
структура	Янковская А. Е.	анализ
прогнозирование	Тетуев Р. К.	Оптимальные
	Пятков М. И.	

Подтема 2

признак	▲ Михайлов Д. В.	▲ Об одном
функционал	Емельянов Г. М.	подходе к синтезу
закономерность	Янковская А. Е.	алгоритмов
даты	Пустовойтов Н.	коррекции
символ	Ю. Федотов Н.	

Фоновая тема 3

предикат	▲ Назипова Н. Н.	▲ Применение
изображение	Чалей М. Б.	генетических
функция	Федотов Н. Г.	алгоритмов в
модель	Тетуев Р. К.	задаче
число	Дедус Ф. Ф.	классификации

Слова(366)

признак
последовательность
строка
днк
скрыть
длина
функционал
периодичность
значение
структура
кодировать
повтор
символ
метод
профильный
закономерность
спектр
тест
генетический
мочь
случайный
коррекция
профильность
прогнозирование
контекст
район
статистический
полуметрика
многочлен
пар
сеть
интерес
тандемный
сая
алфавит
период
паттерн
неприводимый
пользователь
выявление
характеристический
белок
кодирование

Авторы(163)

▲ Чалей М. Б.
Кутыркин В. А.
Назипова Н. Н.
Янковская А. Е.
Тетуев Р. К.
Михайлов Д. В.
Емельянов Г. М.
Дедус Ф. Ф.
Федотов Н. Г.
Пустовойтов Н. Ю.
Панкратов А. Н.
Пятков М. И.
Дюкова Е. В.
Ольшевец М. М.
Сулимова В. В.
Руднев В. Р.
Куликова Л. И.
Романов С. В.
Леухин А. Н.
Мокшанина Д. А.
Хачай М. Ю.
Мирошниченко Л. А.
Гусев В. Д.
Торшин И. Ю.
Шульга Л. А.
Смолякин О. А.
Кольчугин А. С.
Кириллов А. Н.
Ивахненко А. А.
Воронцов К. В.
Туркин П. Ю.
Кузнецов М. Р.
Мекедов И. С.
Сенько О. В.
Середин О. С.
Жданов С. А.
Муравьева О. В.
Романенко А. А.
Соколов А. В.
Киселев М. В.
Дробков
Янковская
Нарикова Л. А.

Статьи(114)

- Скрытая периодичность в кодирующих последовательностях ДНК
- Структурные различия кодирующих и не кодирующих районов ДНК
- Профильно-статистическая основа локальных сигналов в ДНК
- Скрытая профильная периодичность как новый тип периодичности генома
- Оптимизация *
- Интеллектуальный анализ
- Об одном подходе к синтезу алгоритмов коррекции локального возмущения в конечной полуметрике
- Задача снижения размерности в предсказательном моделировании
- Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний
- Характеристики сжатия недоопределенных данных
- Использование сложности разложений в задачах анализа символьных последовательностей
- Поиск схожих пользователей в социальных сетях методами коллаборативной фильтрации
- Методы исследования взаимосвязей в сложных объектах, основанных на сетях закономерностей
- Анализ данных и знаний на основе конвергенции нескольких наук и научных направлений
- Оптимальные байесовские стратегии анализа релевантности для объектов с заданной структурой
- О корректном понижении значности данных в задачах распознавания
- Семантическая схожесть текстов в задаче автоматизированного контроля знаний
- Выявление интересов пользователей социальных сетей методами

Способ построения иерархии

Иерархия строится **рекурсивно**.

ПостроитьУзел(n_{dw}, T):

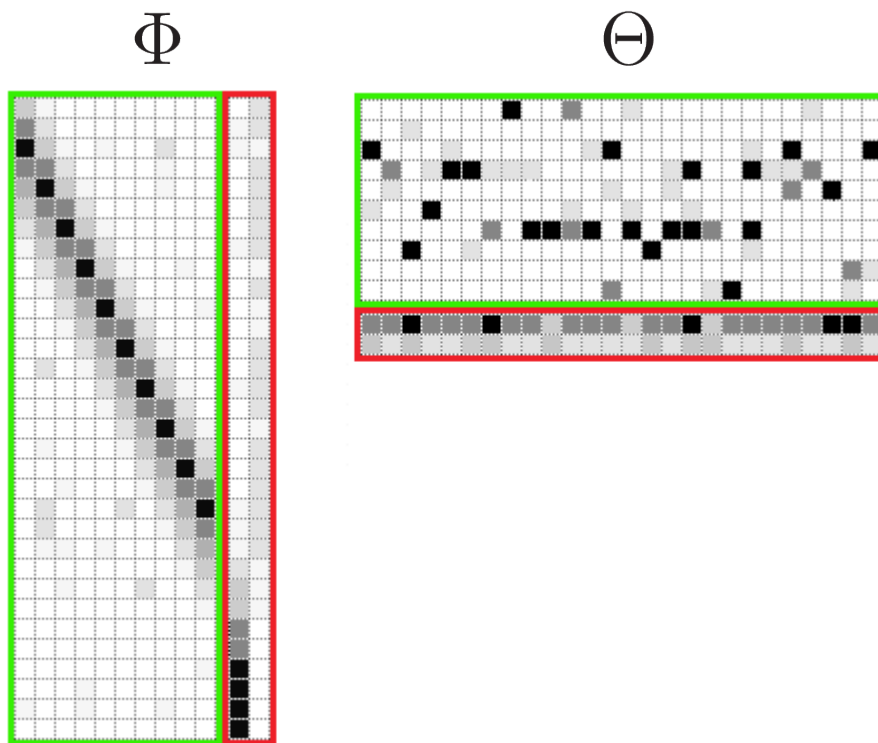
1. Построить плоскую модель — получить матрицы Φ , Θ
2. Разделить входную коллекцию на T коллекций

$$n_{dw} \Rightarrow \{n_{dw}^t\}_{t=1}^T$$

3. для всех $t = 1, \dots, T$
ПостроитьУзел(n_{dw}^t, T_t)

Количество тем T (пока) задается вручную.

Построение плоской модели в узле иерархии



- Выделяем небольшое число предметных тем-детей и несколько фоновых тем
- В некорневых вершинах фоновые темы теперь являются темами общей лексики данного уровня
- Используем базовый набор регуляризаторов:
 - Сглаживание фоновых тем
 - Разреживание предметных тем
 - Декоррелирование

Разделение коллекции на подколлекции

Для каждого слова w в документе d вычисляем $p(t|d, w)$ и делим слово пропорционально между темами

$$p(t|d, w) \propto p(w|t)p(t|d)$$

$$n_{dw}^t = n_{dw}p(t|d, w)$$