Иерархические вероятностные тематические модели

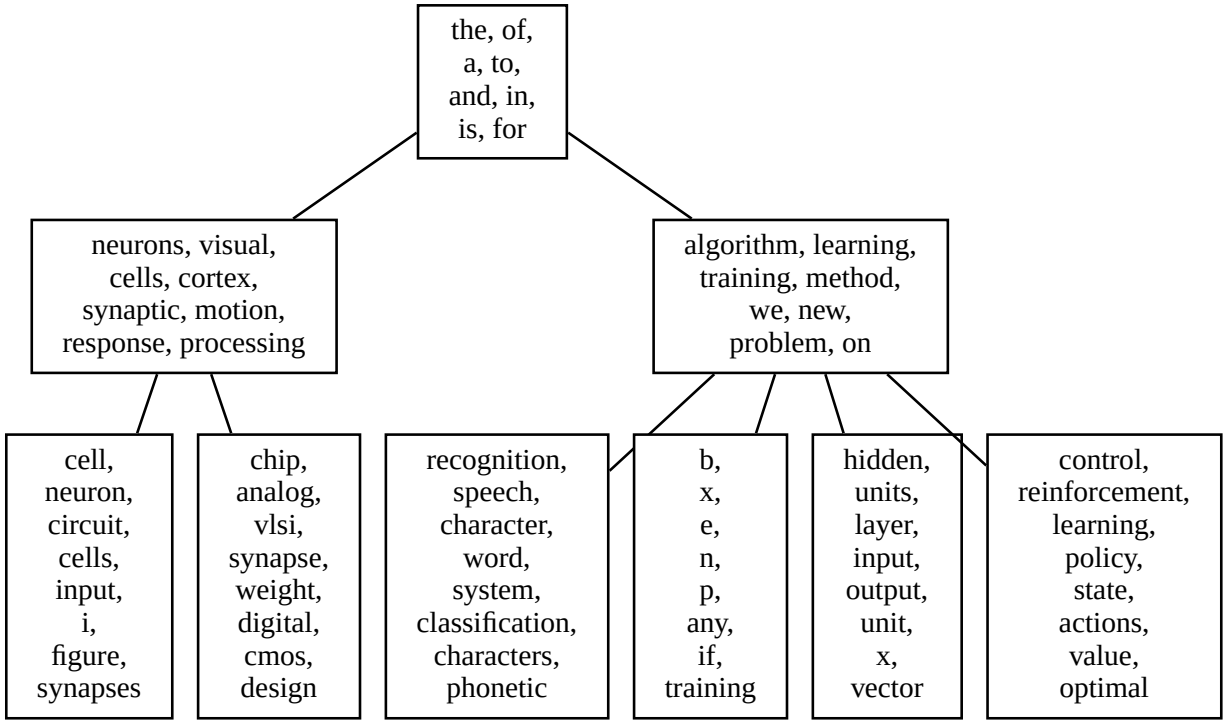
Надежда Чиркова [1]
Научный руководитель: Воронцов К. В. [2]

[1]ВМК МГУ [2]ВЦ РАН

Конференция «Ломоносов 2015», 15.04.2015

План выступления

Иерархическая тематическая модель — это многодольный граф тем. Каждая тема — это набор слов, авторов, документов, ... + множество подтем



✓Тематические иерархии легко интерпретируемы и понятны человеку

✓Иерархическая структура упрощает навигацию по коллекции документов

David M Blei, Hierarchical Topic Models and the Nested Chinese Restaurant Process, 2010

Постановка задачи

D - коллекция текстовых документов

W- множество терминов

Дана коллекция текстовых документов:

n_{dw} - матрица частот слов в документах (мешок слов):

$$F_{dw} = p(w|d) = \frac{n_{dw}}{n_d}$$

Построить модель:

построить L уровней иерархии

Постановка задачи: построение первого уровня

Дано: $F_{dw} = p(w|d) = \frac{n_{dw}}{n_d}$
Построить модель:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td} \Leftrightarrow F = \Phi\Theta$$

с параметрами $\Phi = \{\varphi_{wt}\}_{W \times T}$ и $\Theta = \{\theta_{td}\}_{T \times D}$:
$\varphi_{wt} = p(w|t)$ — распределение слов в теме t;
$\theta_{td} = p(t|d)$ — распределение тем в документе d.
Критерий: максимизация логарифма правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_{t} \varphi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

Постановка задачи: построение нижних уровней иерархии

Дано: $F_{dw} = p(w|d) = \frac{n_{dw}}{n_d}$

Построить модель:

$$p(w|d) = \sum_{t \in T} \sum_{s \in S} p(w|s) p(s|t) p(t|d) = \sum_{t \in T} \sum_{s \in S} \varphi_{ws} \psi_{st} \pi_{td} \Leftrightarrow F = \Phi \Psi \Pi$$

с параметрами $\Phi = \{\varphi_{ws}\}_{W \times T}$, $\Psi = \{\psi_{st}\}_{S \times T}$:

$\varphi_{ws} = p(w|s)$ — распределение слов в дочерней теме s;

$\varphi_{st} = p(s|t)$ — распределение дочерних тем в родительской теме t.

Распределение родительских тем в документе d $\pi_{td} = p(t|d)$ фиксировано (вычисляется на предыдущем уровне)

Критерий: максимизация логарифма правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_{t \in T} \sum_{s \in S} \varphi_{ws} \psi_{st} \pi_{td} + R(\Phi, \Psi, \Pi) \to \max_{\Phi, \Psi},$$

Обучение модели первого уровня

ARTM – регуляризованный PLSA (Probabilistic Latent Semantic Analysis)

Основан на итерационном алгоритме:

EM-алгоритм

E-шаг: $p(t|d,w) = \dfrac{p(w|t)p(t|d)}{\sum_{t' \in T} p(w|t')p(t'|d)} = \dfrac{\varphi_{wt}\theta_{td}}{\sum_{t' \in T} \varphi_{wt'}\theta_{t'd}}$;

M-шаг: $\varphi_{wt} \propto \left(n_{wt} + \varphi_{wt} \dfrac{\partial R}{\partial \varphi_{wt}} \right)_+$; $n_{wt} = \sum_{d \in D} n_{dw} p(t|d,w)$;

$$\theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+ ; \; n_{td} = \sum_{w \in W} n_{dw} p(t|d,w) \; .$$

Обучение моделей нижних уровней

Обобщение ARTM на случай трехматричного разложения
Итерационный алгоритм:

EM-алгоритм

E-шаг: $p(t,s|d,w) = \dfrac{p(w|s)p(s|t)p(t|d)}{\sum_{t'\in T}\sum_{s'\in S} p(w|s')p(s'|t')p(t'|d)} = \dfrac{\varphi_{ws}\psi_{st}\pi_{td}}{\sum_{s'\in S}\sum_{t'\in T}\varphi_{ws'}\psi_{s't'}\pi_{t'd}}$;

M-шаг: $\varphi_{ws} \propto \left(n_{ws} + \varphi_{ws}\dfrac{\partial R}{\partial \varphi_{ws}}\right)_+$; $n_{ws} = \sum_{d\in D}\sum_{t\in T} n_{dw}p(t,s|d,w)$

$\psi_{st} \propto \left(n_{st} + \psi_{st}\dfrac{\partial R}{\partial \psi_{st}}\right)_+$; $n_{st} = \sum_{d\in D}\sum_{w\in W} n_{dw}p(t,s|d,w)$.

Требования к тематическим распределениям слов:

- На каждом уровне выделяем множества предметных и фоновых тем (по одной на каждую предметную тему родительского уровня); в корневой вершине фоновая тема – это слова общей лексики языка и коллекции, на следующих уровнях – слова общей лексики родительских тем;
- Предметные темы декоррелированы между собой и с фоновыми темами;
- Предметные и фоновые темы разрежены, т.е. содержат небольшое количество слов.

Формализация требований в виде регуляризатора матрицы Φ:

$$R(\Phi) = -\tau_1 \sum_{w \in W} \sum_{s \in Sub} \beta_w ln \varphi_{ws} - \tau_2 \sum_{w \in W} \sum_{s \in Bcg} \beta_w ln \varphi_{ws} - \tau_3 \sum_{s,s' \in S} \sum_{w \in W} \varphi_{ws} \varphi_{ws'}$$

Возможные конфигурации матрицы Ψ:

- Если Ψ плотная, то мы имеем полный многодольный граф, если разреженная – граф становится деревом;
- Разреживание Ψ позволяет модели самой распределять потомков между родителями, определять их количество, а также разрешает множественное наследование потомков;
- Обнуляя некоторые значения в начальной инициализации Ψ, мы наоборот задаем структуру графа извне.

Регуляризаторы матрицы Ψ:

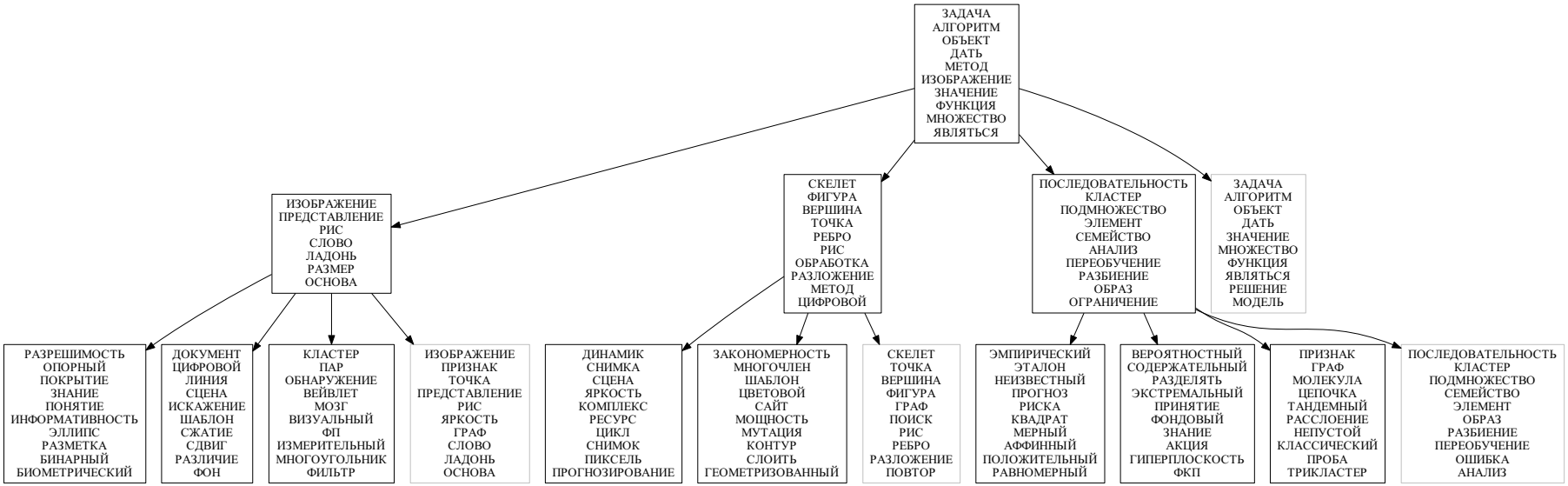
$$R(\Psi) = -\tau_1 \sum_{s \in S} \sum_{t \in T} \ln \psi_{st} + \tau_2 \sum_{s \in Bcg} \sum_{t \in T} \ln \psi_{st}$$

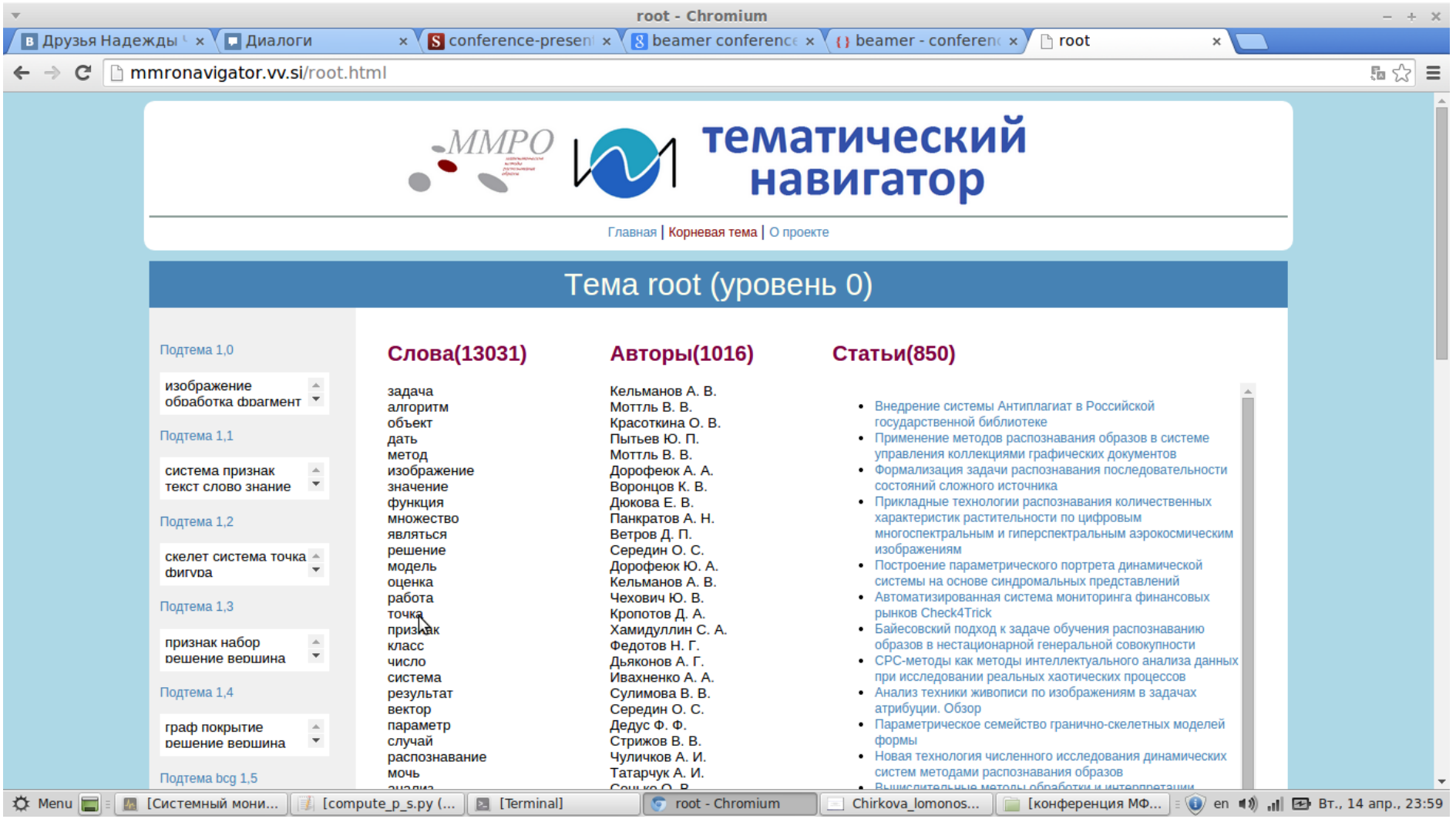
Эксперименты

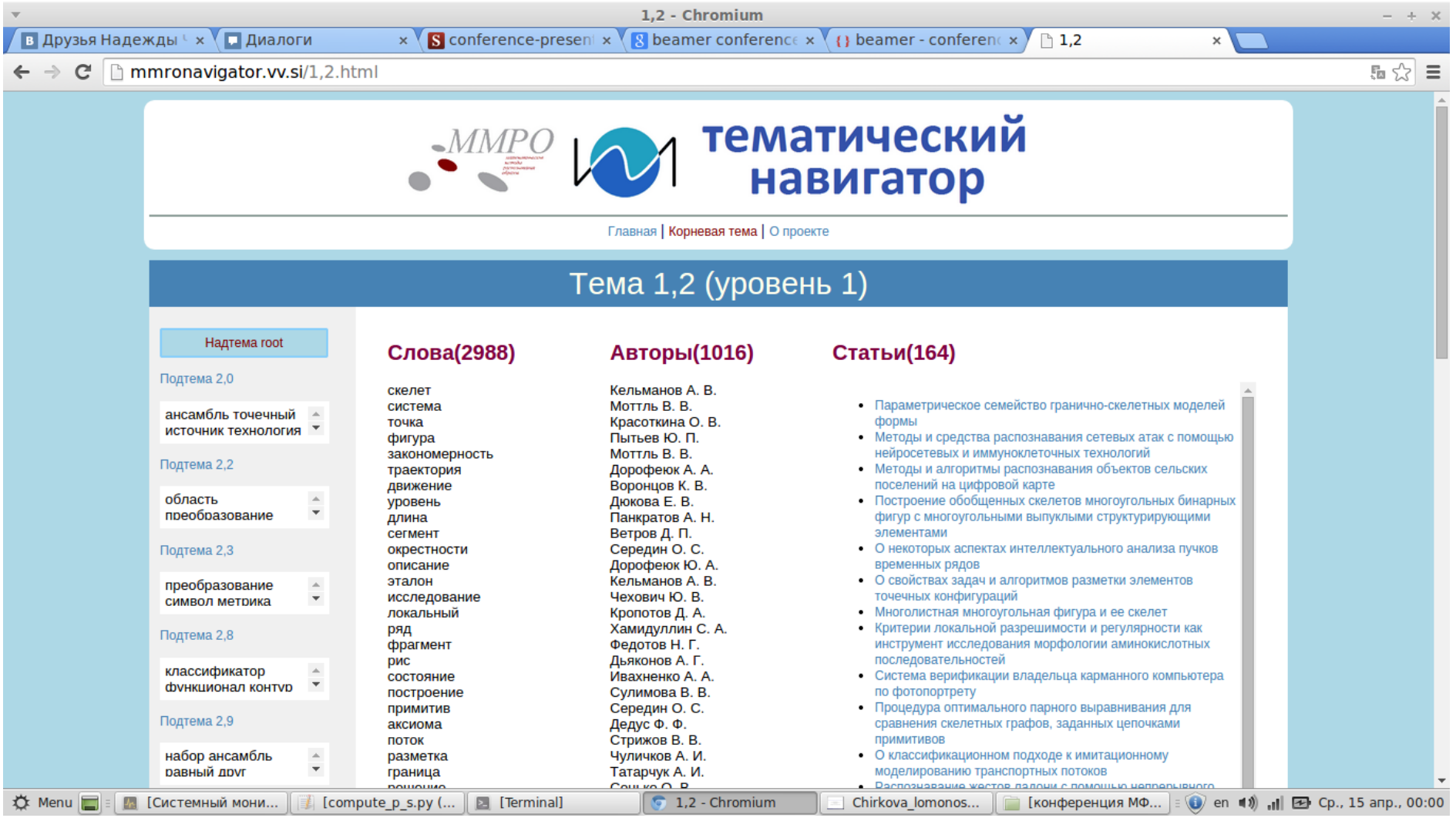
Эксперименты проводятся на коллекции документов конференций «Математические методы распознавания образов» и «Интеллектуализация обработки информации».
$|D| = 850$, $|W| = 13031$
Веб-навигатор:

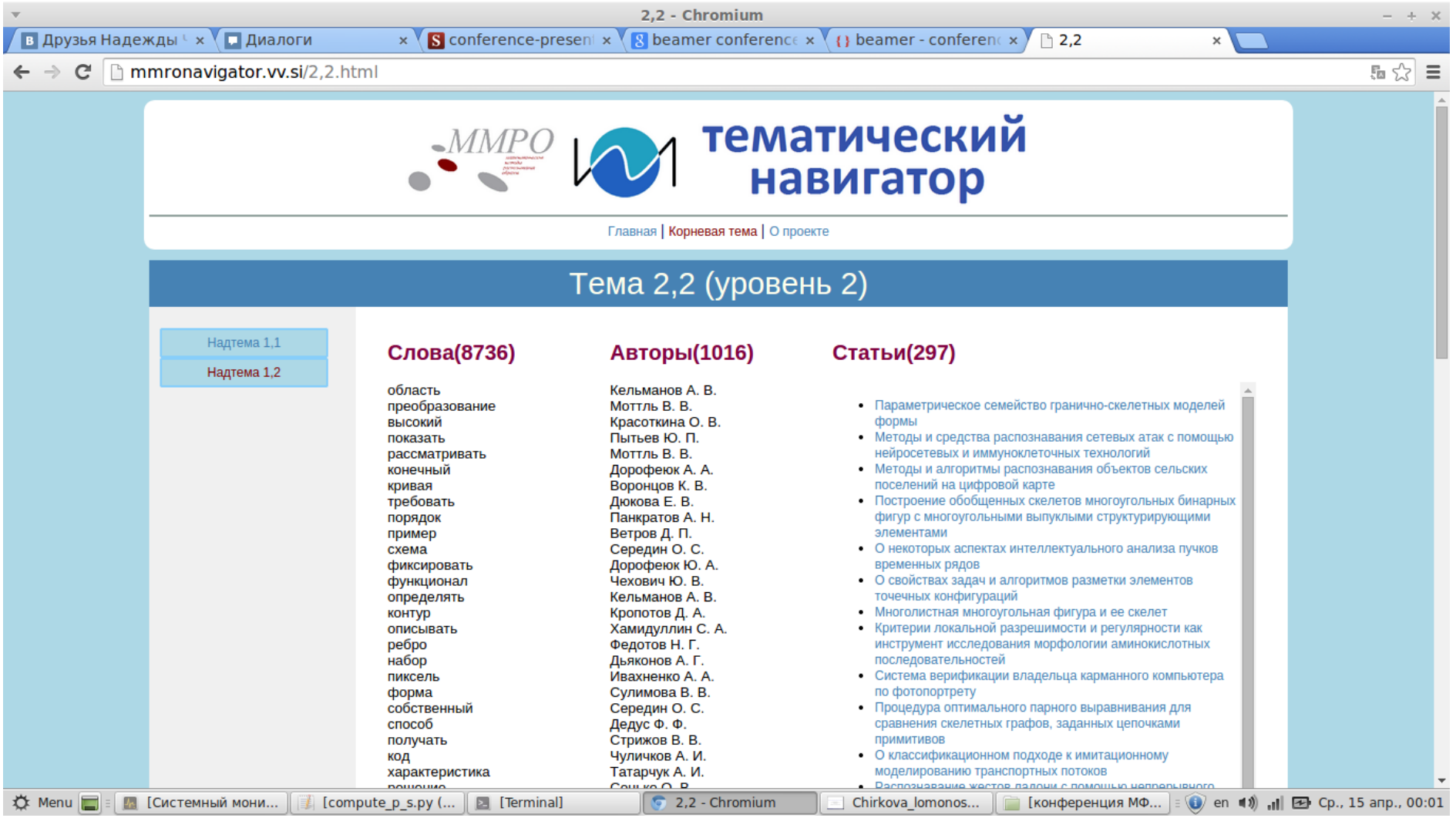
MMROnavigator.vv.si

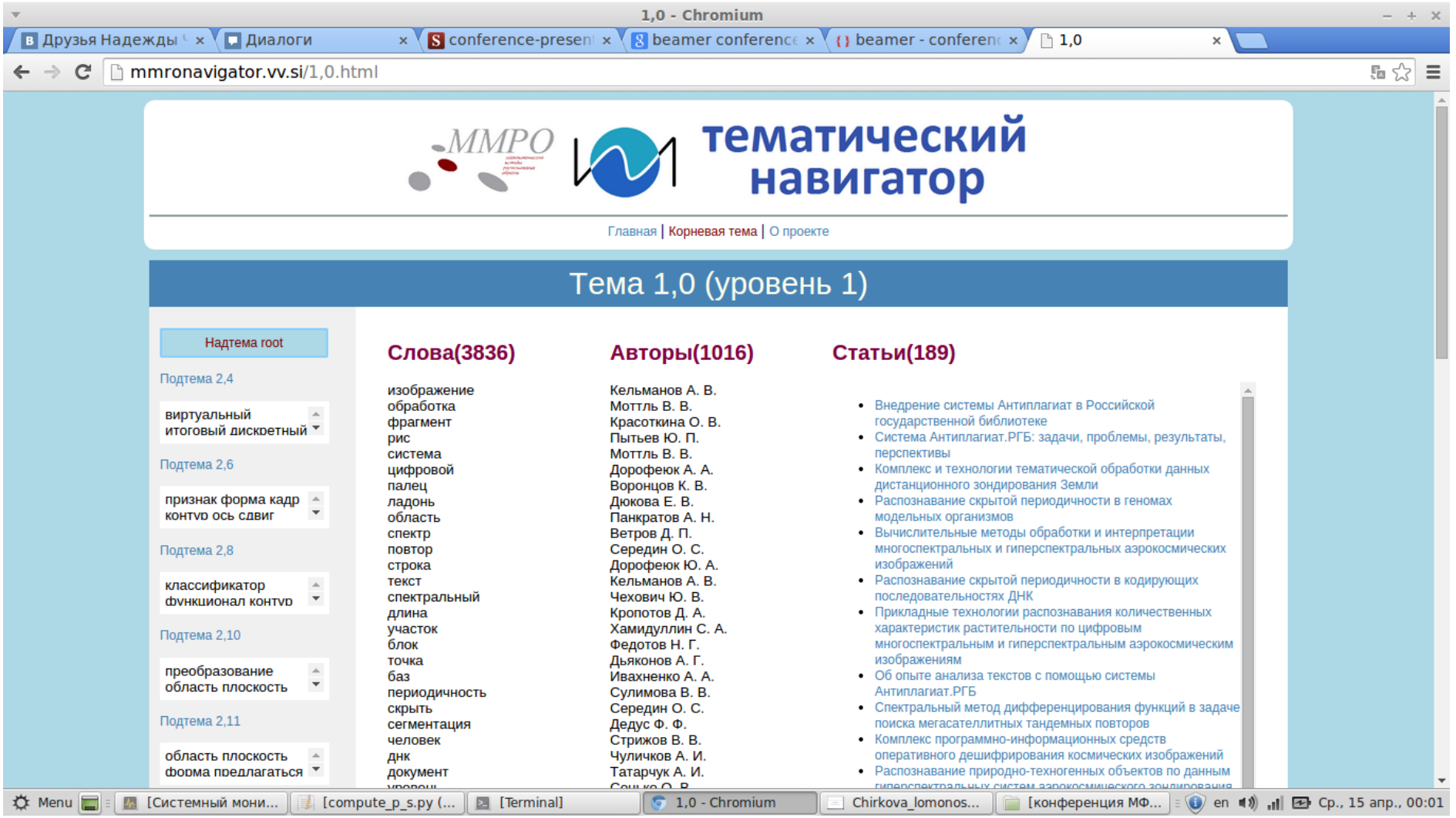
Фрагмент иерархии

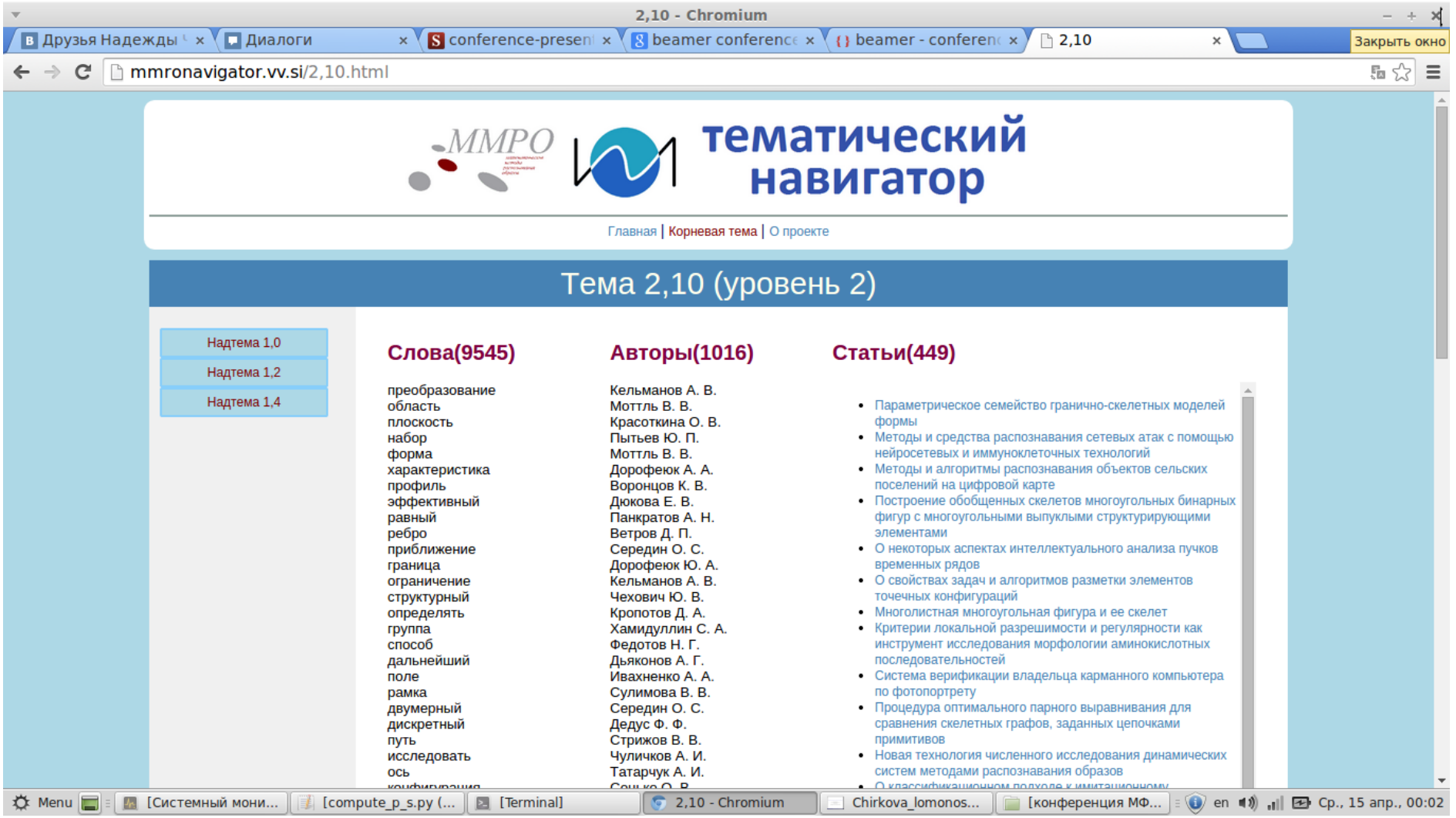



root - Chromium
Друзья Надежды
Диалоги
conference-presen
beamer conferenc
beamer - conferen
root
mmronavigator.vv.si/root.html
ММРО
тематический навигатор
Главная | Корневая тема | О проекте
Тема root (уровень 0)
Подтема 1,0
изображение обработка фрагмент
Подтема 1,1
система признак текст слово знание
Подтема 1,2
скелет система точка фигура
Подтема 1,3
признак набор решение вершина
Подтема 1,4
граф покрытие решение вершина
Подтема bcg 1,5
Слова(13031)
задача
алгоритм
объект
дать
метод
изображение
значение
функция
множество
являться
решение
модель
оценка
работа
точка
признак
класс
число
система
результат
вектор
параметр
случай
распознавание
мочь
Авторы(1016)
Кельманов А. В.
Моттль В. В.
Красоткина О. В.
Пытьев Ю. П.
Моттль В. В.
Дорофеюк А. А.
Воронцов К. В.
Дюкова Е. В.
Панкратов А. Н.
Ветров Д. П.
Середин О. С.
Дорофеюк Ю. А.
Кельманов А. В.
Чехович Ю. В.
Кропотов Д. А.
Хамидуллин С. А.
Федотов Н. Г.
Дьяконов А. Г.
Ивахненко А. А.
Сулимова В. В.
Середин О. С.
Дедус Ф. Ф.
Стрижов В. В.
Чуличков А. И.
Татарчук А. И.
Статьи(850)
Внедрение системы Антиплагиат в Российской государственной библиотеке
Применение методов распознавания образов в системе управления коллекциями графических документов
Формализация задачи распознавания последовательности состояний сложного источника
Прикладные технологии распознавания количественных характеристик растительности по цифровым многоспектральным и гиперспектральным аэрокосмическим изображениям
Построение параметрического портрета динамической системы на основе синдромальных представлений
Автоматизированная система мониторинга финансовых рынков Check4Trick
Байесовский подход к задаче обучения распознаванию образов в нестационарной генеральной совокупности
CPC-методы как методы интеллектуального анализа данных при исследовании реальных хаотических процессов
Анализ техники живописи по изображениям в задачах атрибуции. Обзор
Параметрическое семейство гранично-скелетных моделей формы
Новая технология численного исследования динамических систем методами распознавания образов
Menu
[Системный мони...
[compute_p_s.py (...
[Terminal]
root - Chromium
Chirkova_lomonos...
[конференция МФ...
en
Вт., 14 апр., 23:59


1,2 - Chromium
Друзья Надежды
Диалоги
conference-presen
beamer conferenc
beamer - conferen
1,2
mmronavigator.vv.si/1,2.html
ММРО
тематический навигатор
Главная | Корневая тема | О проекте
Тема 1,2 (уровень 1)
Надтема root
Подтема 2,0
ансамбль точечный источник технология
Подтема 2,2
область преобразование
Подтема 2,3
преобразование символ метрика
Подтема 2,8
классификатор функционал контур
Подтема 2,9
набор ансамбль равный двух
Слова(2988)
скелет
система
точка
фигура
закономерность
траектория
движение
уровень
длина
сегмент
окрестности
описание
эталон
исследование
локальный
ряд
фрагмент
рис
состояние
построение
примитив
аксиома
поток
разметка
граница
Авторы(1016)
Кельманов А. В.
Моттль В. В.
Красоткина О. В.
Пытьев Ю. П.
Моттль В. В.
Дорофеюк А. А.
Воронцов К. В.
Дюкова Е. В.
Панкратов А. Н.
Ветров Д. П.
Середин О. С.
Дорофеюк Ю. А.
Кельманов А. В.
Чехович Ю. В.
Кропотов Д. А.
Хамидуллин С. А.
Федотов Н. Г.
Дьяконов А. Г.
Ивахненко А. А.
Сулимова В. В.
Середин О. С.
Дедус Ф. Ф.
Стрижов В. В.
Чуличков А. И.
Татарчук А. И.
Статьи(164)
Параметрическое семейство гранично-скелетных моделей формы
Методы и средства распознавания сетевых атак с помощью нейросетевых и иммуноклеточных технологий
Методы и алгоритмы распознавания объектов сельских поселений на цифровой карте
Построение обобщенных скелетов многоугольных бинарных фигур с многоугольными выпуклыми структурирующими элементами
О некоторых аспектах интеллектуального анализа пучков временных рядов
О свойствах задач и алгоритмов разметки элементов точечных конфигураций
Многолистная многоугольная фигура и ее скелет
Критерии локальной разрешимости и регулярности как инструмент исследования морфологии аминокислотных последовательностей
Система верификации владельца карманного компьютера по фотопортрету
Процедура оптимального парного выравнивания для сравнения скелетных графов, заданных цепочками примитивов
О классификационном подходе к имитационному моделированию транспортных потоков
Menu
[Системный мони...
[compute_p_s.py (...
[Terminal]
1,2 - Chromium
Chirkova_lomonos...
[конференция МФ...
en
Ср., 15 апр., 00:00


2,2 - Chromium
Друзья Надежды
Диалоги
conference-presen
beamer conferenc
beamer - conferen
2,2
mmronavigator.vv.si/2,2.html
ММРО
тематический навигатор
Главная | Корневая тема | О проекте
Тема 2,2 (уровень 2)
Надтема 1,1
Надтема 1,2
Слова(8736)
область
преобразование
высокий
показать
рассматривать
конечный
кривая
требовать
порядок
пример
схема
фиксировать
функционал
определять
контур
описывать
ребро
набор
пиксель
форма
собственный
способ
получать
код
характеристика
Авторы(1016)
Кельманов А. В.
Моттль В. В.
Красоткина О. В.
Пытьев Ю. П.
Моттль В. В.
Дорофеюк А. А.
Воронцов К. В.
Дюкова Е. В.
Панкратов А. Н.
Ветров Д. П.
Середин О. С.
Дорофеюк Ю. А.
Чехович Ю. В.
Кельманов А. В.
Кропотов Д. А.
Хамидуллин С. А.
Федотов Н. Г.
Дьяконов А. Г.
Ивахненко А. А.
Сулимова В. В.
Середин О. С.
Дедус Ф. Ф.
Стрижов В. В.
Чуличков А. И.
Татарчук А. И.
Статьи(297)
Параметрическое семейство гранично-скелетных моделей формы
Методы и средства распознавания сетевых атак с помощью нейросетевых и иммуноклеточных технологий
Методы и алгоритмы распознавания объектов сельских поселений на цифровой карте
Построение обобщенных скелетов многоугольных бинарных фигур с многоугольными выпуклыми структурирующими элементами
О некоторых аспектах интеллектуального анализа пучков временных рядов
О свойствах задач и алгоритмов разметки элементов точечных конфигураций
Многолистная многоугольная фигура и ее скелет
Критерии локальной разрешимости и регулярности как инструмент исследования морфологии аминокислотных последовательностей
Система верификации владельца карманного компьютера по фотопортрету
Процедура оптимального парного выравнивания для сравнения скелетных графов, заданных цепочками примитивов
О классификационном подходе к имитационному моделированию транспортных потоков
Menu
[Системный мони...
[compute_p_s.py (...
[Terminal]
2,2 - Chromium
Chirkova_lomonos...
[конференция МФ...
en
Ср., 15 апр., 00:01


1,0 - Chromium
Друзья Надежды
Диалоги
conference-presen
beamer conferenc
beamer - conferen
1,0
mmronavigator.vv.si/1,0.html
ММРО
тематический навигатор
Главная | Корневая тема | О проекте
Тема 1,0 (уровень 1)
Надтема root
Подтема 2,4
виртуальный итоговый дискретный
Подтема 2,6
признак форма кадр контур ось сдвиг
Подтема 2,8
классификатор функционал контур
Подтема 2,10
преобразование область плоскость
Подтема 2,11
область плоскость форма предлагаться
Слова(3836)
изображение
обработка
фрагмент
рис
система
цифровой
палец
ладонь
область
спектр
повтор
строка
текст
спектральный
длина
участок
блок
точка
баз
периодичность
скрыть
сегментация
человек
днк
документ
Авторы(1016)
Кельманов А. В.
Моттль В. В.
Красоткина О. В.
Пытьев Ю. П.
Моттль В. В.
Дорофеюк А. А.
Воронцов К. В.
Дюкова Е. В.
Панкратов А. Н.
Ветров Д. П.
Середин О. С.
Дорофеюк Ю. А.
Кельманов А. В.
Чехович Ю. В.
Кропотов Д. А.
Хамидуллин С. А.
Федотов Н. Г.
Дьяконов А. Г.
Ивахненко А. А.
Сулимова В. В.
Середин О. С.
Дедус Ф. Ф.
Стрижов В. В.
Чуличков А. И.
Татарчук А. И.
Статьи(189)
Внедрение системы Антиплагиат в Российской государственной библиотеке
Система Антиплагиат.РГБ: задачи, проблемы, результаты, перспективы
Комплекс и технологии тематической обработки данных дистанционного зондирования Земли
Распознавание скрытой периодичности в геномах модельных организмов
Вычислительные методы обработки и интерпретации многоспектральных и гиперспектральных аэрокосмических изображений
Распознавание скрытой периодичности в кодирующих последовательностях ДНК
Прикладные технологии распознавания количественных характеристик растительности по цифровым многоспектральным и гиперспектральным аэрокосмическим изображениям
Об опыте анализа текстов с помощью системы Антиплагиат.РГБ
Спектральный метод дифференцирования функций в задаче поиска мегасателлитных тандемных повторов
Комплекс программно-информационных средств оперативного дешифрирования космических изображений
Распознавание природно-техногенных объектов по данным
Menu
[Системный мони...
[compute_p_s.py (...
[Terminal]
1,0 - Chromium
Chirkova_lomonos...
[конференция МФ...
en
Ср., 15 апр., 00:01


2,10 - Chromium
Друзья Надежды
Диалоги
conference-presen
beamer conferenc
beamer - conferen
2,10
Закрыть окно
mmronavigator.vv.si/2,10.html
ММРО
тематический навигатор
Главная | Корневая тема | О проекте
Тема 2,10 (уровень 2)
Надтема 1,0
Надтема 1,2
Надтема 1,4
Слова(9545)
преобразование
область
плоскость
набор
форма
характеристика
профиль
эффективный
равный
ребро
приближение
граница
ограничение
структурный
определять
группа
способ
дальнейший
поле
рамка
двумерный
дискретный
путь
исследовать
ось
Авторы(1016)
Кельманов А. В.
Моттль В. В.
Красоткина О. В.
Пытьев Ю. П.
Моттль В. В.
Дорофеюк А. А.
Воронцов К. В.
Дюкова Е. В.
Панкратов А. Н.
Ветров Д. П.
Середин О. С.
Дорофеюк Ю. А.
Кельманов А. В.
Чехович Ю. В.
Кропотов Д. А.
Хамидуллин С. А.
Федотов Н. Г.
Дьяконов А. Г.
Ивахненко А. А.
Сулимова В. В.
Середин О. С.
Дедус Ф. Ф.
Стрижов В. В.
Чуличков А. И.
Татарчук А. И.
Статьи(449)
Параметрическое семейство гранично-скелетных моделей формы
Методы и средства распознавания сетевых атак с помощью нейросетевых и иммуноклеточных технологий
Методы и алгоритмы распознавания объектов сельских поселений на цифровой карте
Построение обобщенных скелетов многоугольных бинарных фигур с многоугольными выпуклыми структурирующими элементами
О некоторых аспектах интеллектуального анализа пучков временных рядов
О свойствах задач и алгоритмов разметки элементов точечных конфигураций
Многолистная многоугольная фигура и ее скелет
Критерии локальной разрешимости и регулярности как инструмент исследования морфологии аминокислотных последовательностей
Система верификации владельца карманного компьютера по фотопортрету
Процедура оптимального парного выравнивания для сравнения скелетных графов, заданных цепочками примитивов
Новая технология численного исследования динамических систем методами распознавания образов
Menu
[Системный мони...
[compute_p_s.py (...
[Terminal]
2,10 - Chromium
Chirkova_lomonos...
[конференция МФ...
en
Ср., 15 апр., 00:02

Результаты и перспективы

Результаты:

- предложен способ построения тематических иерархий с помощью двух- и трехматричных разложений матрицы частот слов;
- разработана пилотная версия тематического навигатора ММРО-ИОИ.

Дальнейшие исследования:

- оценивание и улучшение качества иерархии;
- более детальное исследование влияния различных стратегий регуляризации на качество иерерхии;
- частичное обучение (использование экспертной разметки);
- автоматическое именование тем.

Zavitsanos E., Paliouras G., Vouros G. A. Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // Journal of Machine Learning Research, 2011. Vol. 12. Pp. 2749–2775.

Wang C., Danilevsky M., Desai N., Zhang Y., Nguyen P., Taula T., Han J. A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy // KDD '13, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013. Pp. 437–445.

Wang C., Liu X., Song Y., Han J. Scalable and robust construction of topical hierarchies // arXiv:1403.3460v1 [cs.LG] 13 Mar 2014.

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'2014, Analysis of Images, Social networks and Texts. Springer International Publishing Switzerland. Communications in Computer and Information Science (CCIS). 2014, Vol. 436. Pp. 29–46.