Визуализация результатов картирования групп речевых маркеров

Вдовина Е. А.

Московский физико-технический институт

Научный руководитель: к.ф-м.н. Майсурадзе Арчил Ивериевич

2014

Об исследовании

- Цель — разработка метода визуализации реляционных данных.
- Предметная область для иллюстраций — наукометрия.
- Исходные данные описываются трехдольной полужесткой моделью.
- Данные нужно визуализировать в виде карты.

Задача визуализации

Общая задача визуализации

По исходным данным требуется построить диаграмму заданного *типа*, удовлетворяющую заданному *набору требований*.

Для постановки задачи требуется определить:

- исходные данные;
- тип диаграммы;
- набор требований.

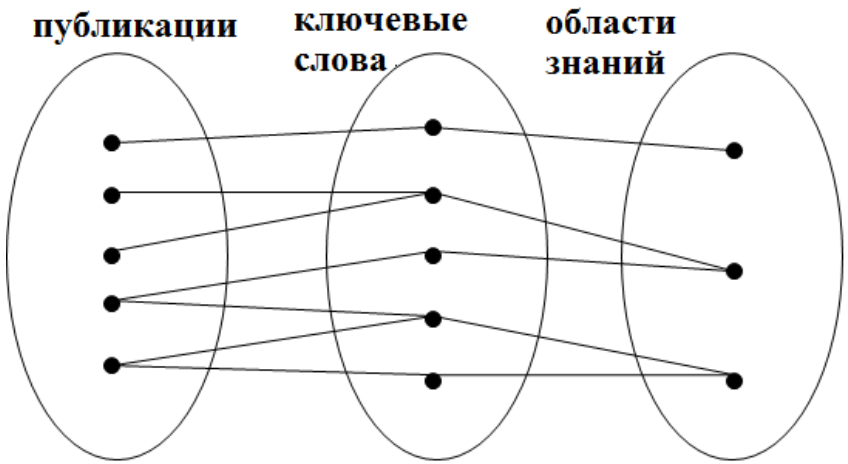
Методология исследования

Этапы классической методологии разработки метода визуализации:

1. Описание данных.
2. Сбор требований у экспертов.
3. Математическая формализация требований.
4. Решение математической задачи.
5. Вычислительный эксперимент.

Выполнены все этапы.

Трехдольная полужесткая модель



Реляционные данные:

- три единицы анализа — объекты (публикации), маркеры (ключевые слова), классы (области знаний);
- отношения между единицами анализа:
  1. публикации – ключевые слова (многие ко многим);
  2. ключевые слова – области знаний (многие к одному);
  3. публикации – области знаний (композиция предыдущих);
- *полужесткая* — есть отношение «многие к одному»;
- отношения только гетерогенные, внутри единиц анализа структуры нет («плоская» структура).

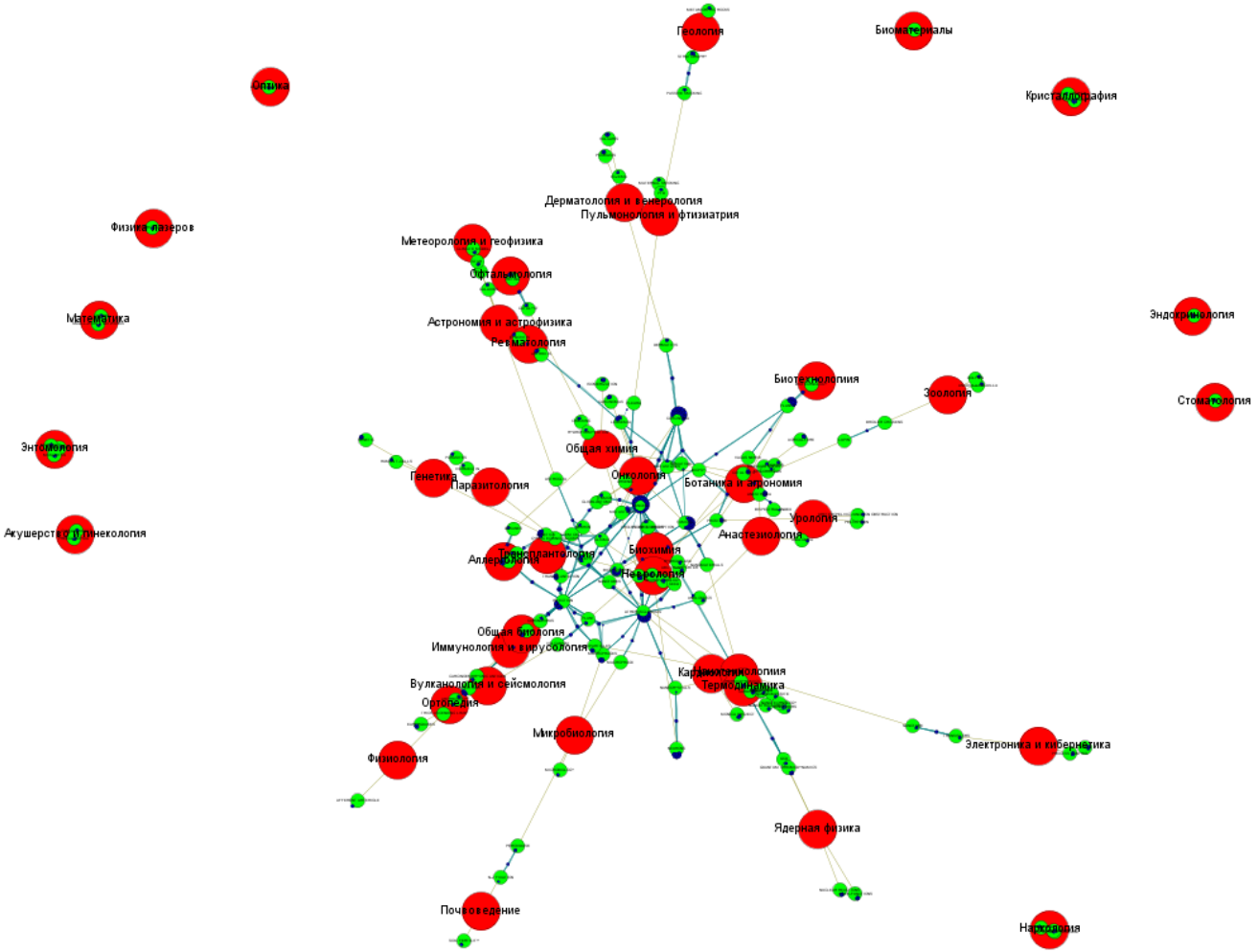
Исходные данные

Результаты иллюстрируются на данных из области наукометрии.

- 42 области знаний (классы);
- 133 ключевых слова (маркеры);
- 1756 публикаций (объекты).

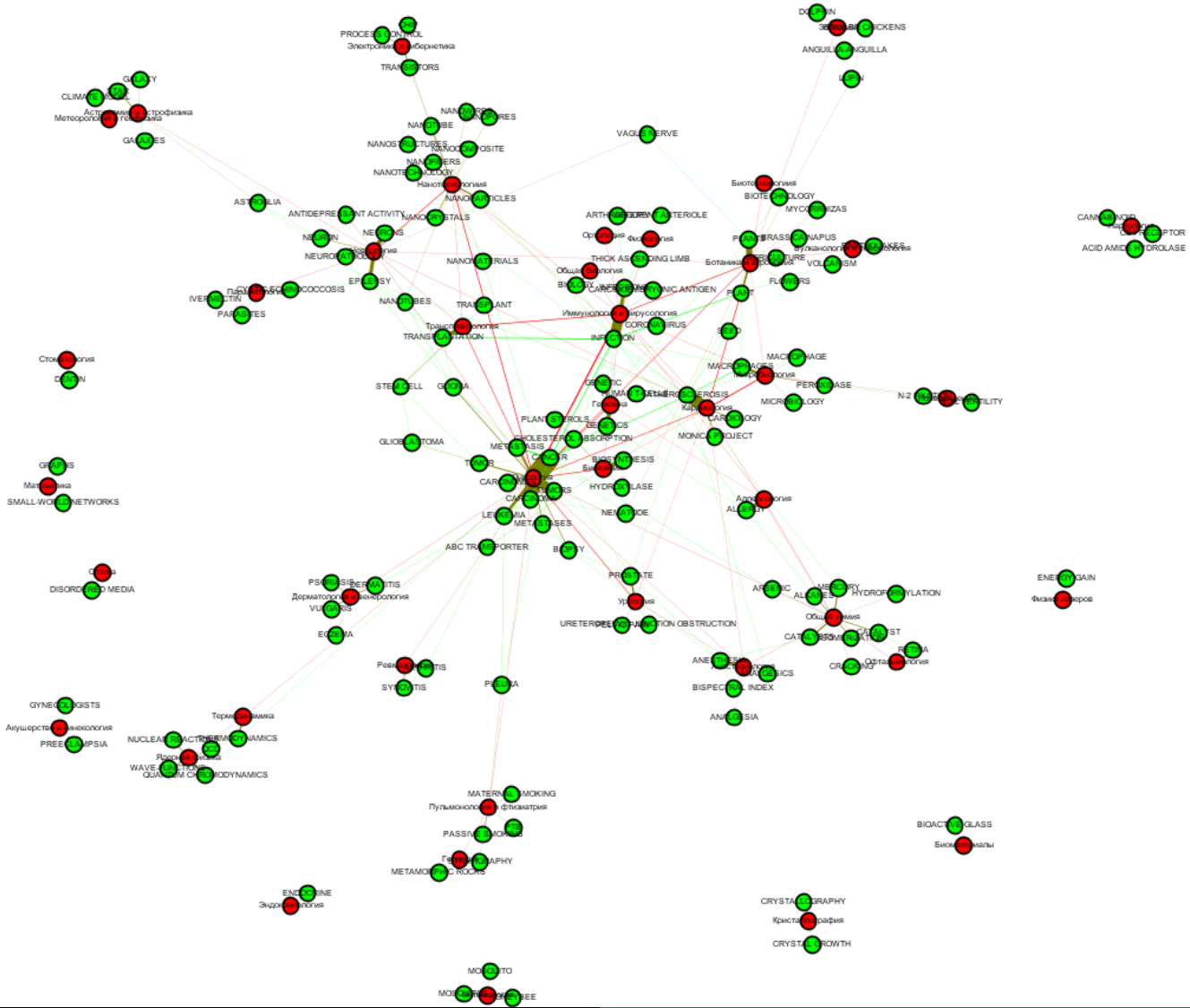
Визуализация графа

Визуализация непосредственно исходных данных. Красные, зеленые и синие вершины — области знаний, ключевые слова и публикации соответственно.



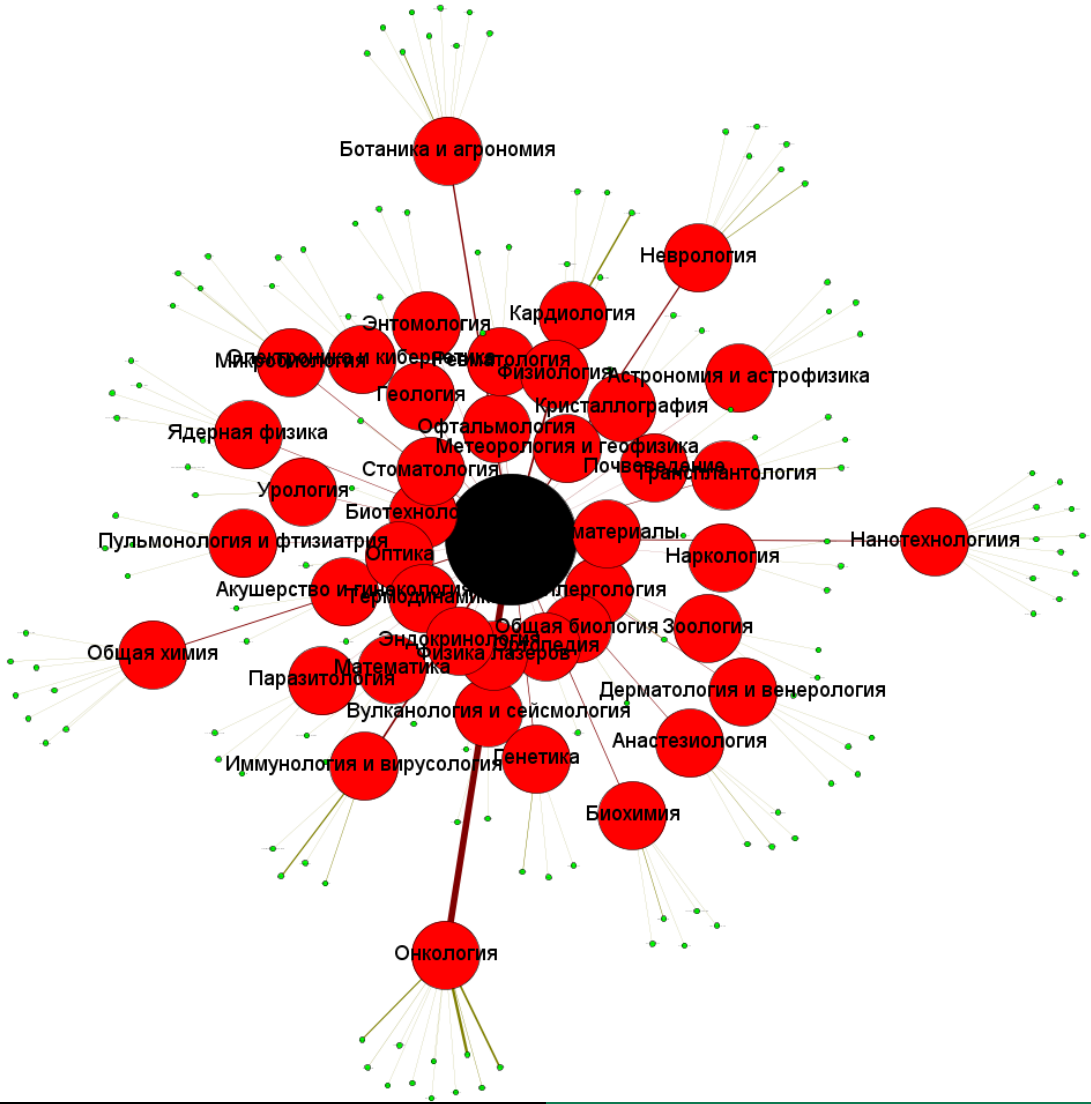
Агрегирование информации

Удаляются публикации, добавляются гомогенные связи, у рёбер появляются веса.



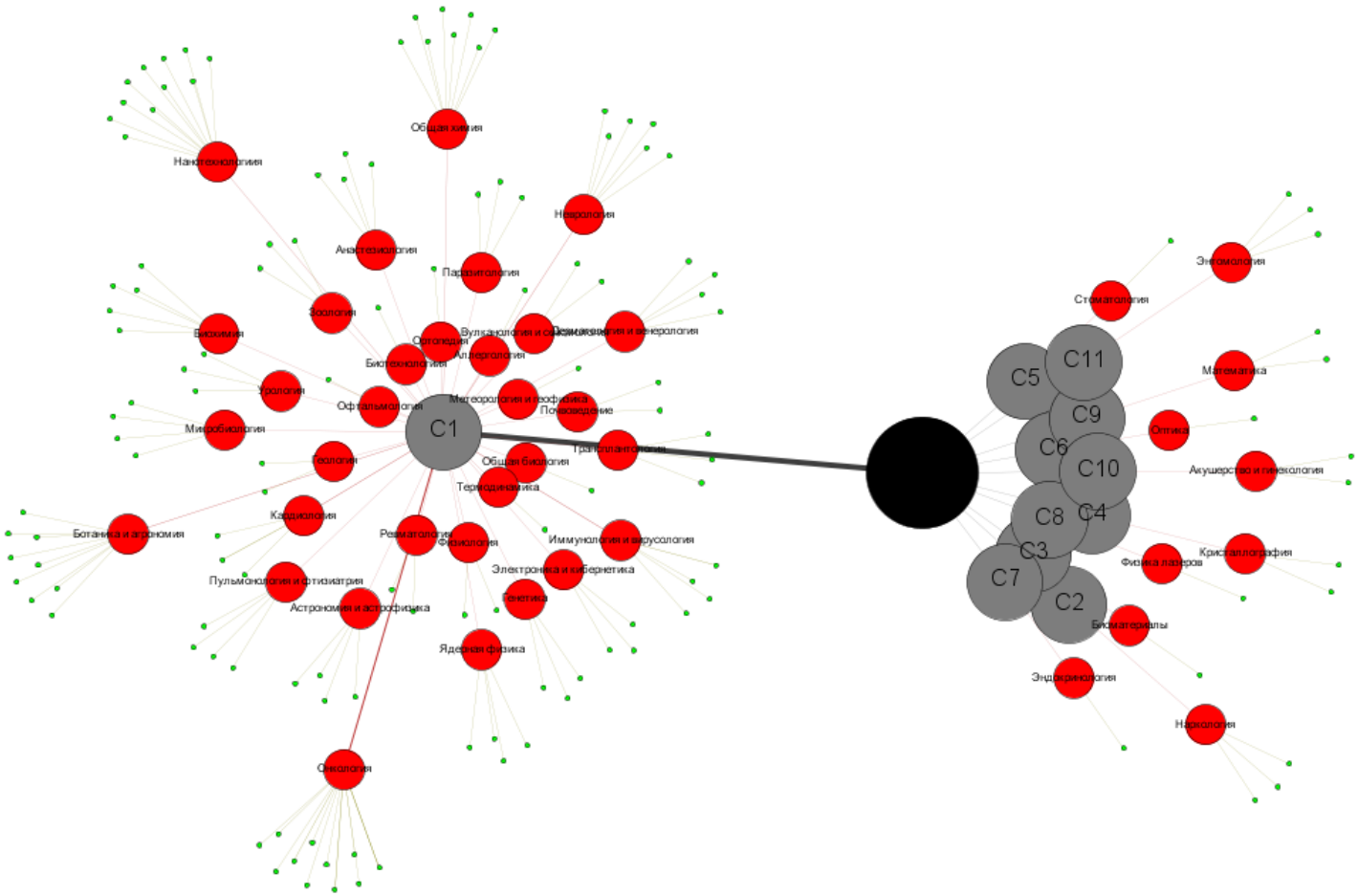
Дерево

Удаляются публикации, добавляется корень (большая черная вершина), все ребра гетерогенные и у них появляются веса.



Еще один уровень иерархии в дереве

Компоненты связности (серые вершины) графа добавляют в дерево ещё один уровень иерархии.



Карта как тип диаграммы

Определение

Единицы отображения — это геометрические объекты, из которых состоит диаграмма.

Определение

Карта — это диаграмма, на которой есть следующие единицы отображения:

- непересекающиеся площадные формы,
- точечные маркеры.

Определение

Площадная форма — это подмножество области построения диаграммы, обычно являющееся односвязной областью. Характеризуется положением, площадью, длиной границы.

Точечные маркеры характеризуются положением.

Требования к карте

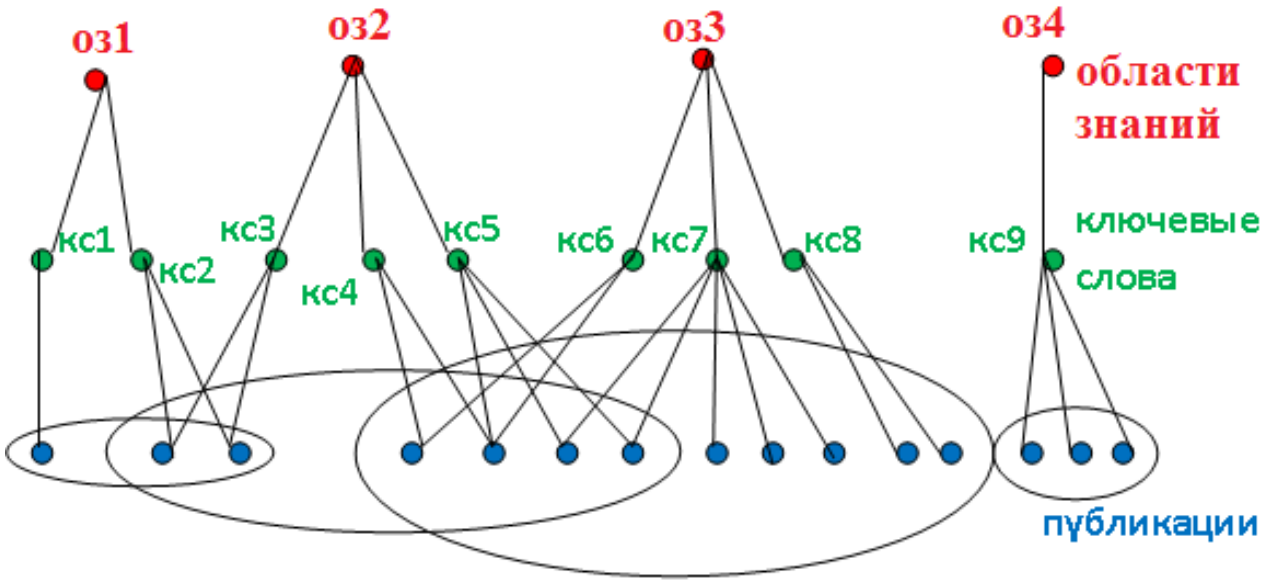
Область построения карты — прямоугольник.
Связь единиц анализа с единицами отображения:

- публикациям ничего не соответствует;
- каждой области знаний соответствует одна форма;
- каждому ключевому слову соответствует один точечный маркер.

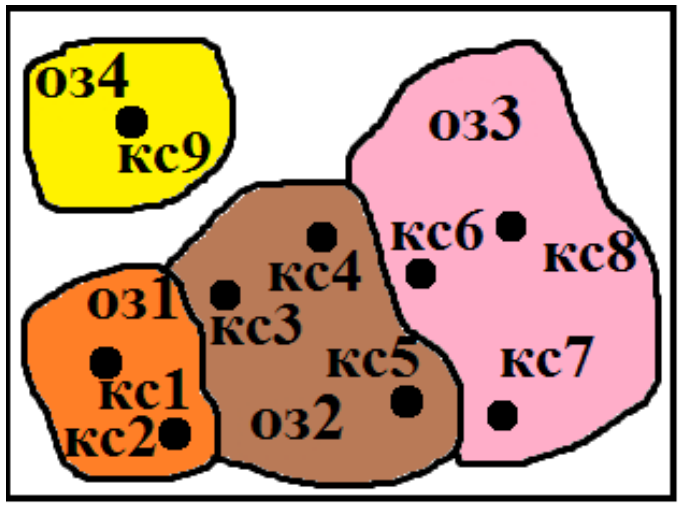
Кроме того:

- на карте могут быть «моря» — области, не входящие ни в одну форму;
- формы являются односвязными.

Требования к карте



Здесь кс2, кс3, кс4, кс5, кс6, кс7 — «пограничные» ключевые слова. Карта для этих данных должна выглядеть примерно так:



Подход к построению карты

Единственное требование к «внутренним» ключевым словам — находиться в «своей» площадной форме. Значит, до самого последнего момента их можно не учитывать. Карта строится в 3 этапа:

1. В исходном графе информация о публикациях агрегируется. Полученный граф делится на компоненты связности и далее каждая рассматривается отдельно.
2. Для каждой компоненты связности решается задача расстановки вершин графа.
3. Вершины графа используются в качестве сайтов для диаграммы Вороного. Одна форма — объединение ячеек, соответствующих области знаний и ее «пограничным» ключевым словам.

Расстановка вершин графа

Подход к расположению вершин:

- приближенно рассмотреть расположение форм как расположение кругов на плоскости;
- уточнить границы форм с помощью расположения «пограничных» ключевых слов.

Для формализации требований к положению вершин и решения математической задачи используется force-directed подход.

Решение

$\mathbf{x}_i, \mathbf{y}_j$ — координаты i-ого ключевого слова и j-ой области знаний; $\rho(\mathbf{a}, \mathbf{b})$ — евклидова метрика.

1. оптимизационная задача для координат областей знаний:

$$TE = \sum_{i,j} \chi(\mathbf{y}_i, \mathbf{y}_j) \to \min,$$

$$\chi(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} M\left(\rho(\mathbf{y}_i, \mathbf{y}_j) - (R_i + R_j)\right)^2, \text{формы граничат;} \\ 0, \text{формы не граничат и} \rho(\mathbf{y}_i, \mathbf{y}_j) \geqslant R_i + R_j \\ L\left(\rho(\mathbf{y}_i, \mathbf{y}_j) - (R_i + R_j)\right)^2, \text{ иначе} \end{cases}$$

где L, M — некоторые коэффициенты.

2. координаты областей знаний использовать как начальное приближение для задачи

$$W\sum_{k,i} f(\mathbf{y}_k, \mathbf{x}_i) + \sum_{i,j}\left(Hg_{ij}\rho(\mathbf{x}_i, \mathbf{x}_j)^2 + G\frac{r_i + r_j}{\alpha\rho(\mathbf{x}_i, \mathbf{x}_j)}\right) +$$

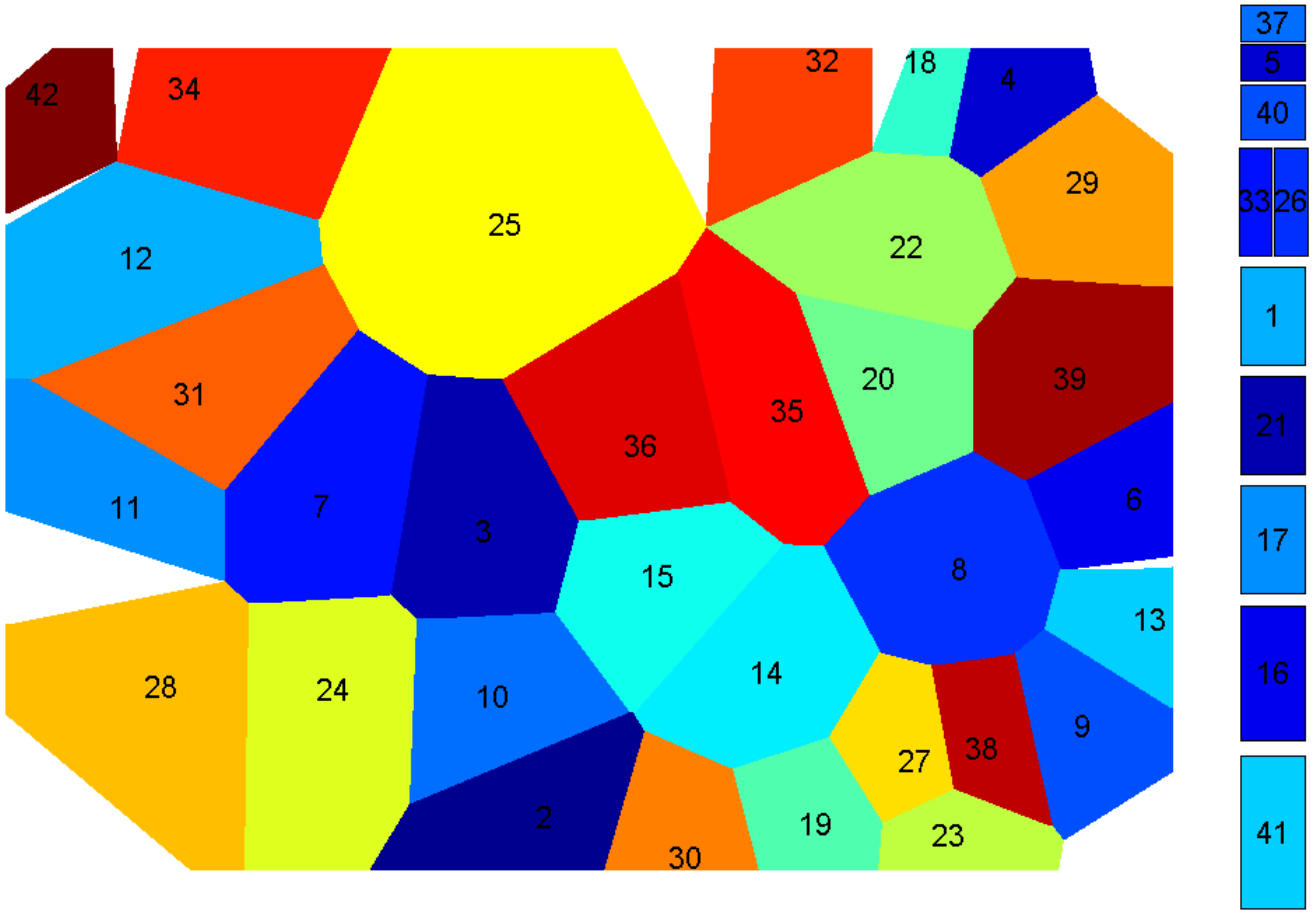
$$+V\sum_{i,j}\frac{R_i + R_j}{\alpha\rho(\mathbf{y}_i, \mathbf{y}_j)} \to \min,$$

$$f(\mathbf{y}_k, \mathbf{x}_i) = \begin{cases} \left(\frac{\rho(\mathbf{y}_k, \mathbf{x}_i)}{R_k}\right)^8 - 1, \text{ключевое слово } i \\ \text{относится к области знаний } k \text{ и} \rho(\mathbf{y}_k, \mathbf{x}_i) > R_k; \\ -\ln\frac{\rho(\mathbf{y}_k, \mathbf{x}_i)}{R_k}, \text{ключевое слово } i \\ \text{не относится к области знаний } k \text{ и} \rho(\mathbf{y}_k, \mathbf{x}_i) < R_k; \\ 0, \text{ иначе} \end{cases}$$

где W, H, G, V — некоторые коэффициенты.

Промежуточный этап: координаты областей знаний, полученные без учета координат ключевых слов

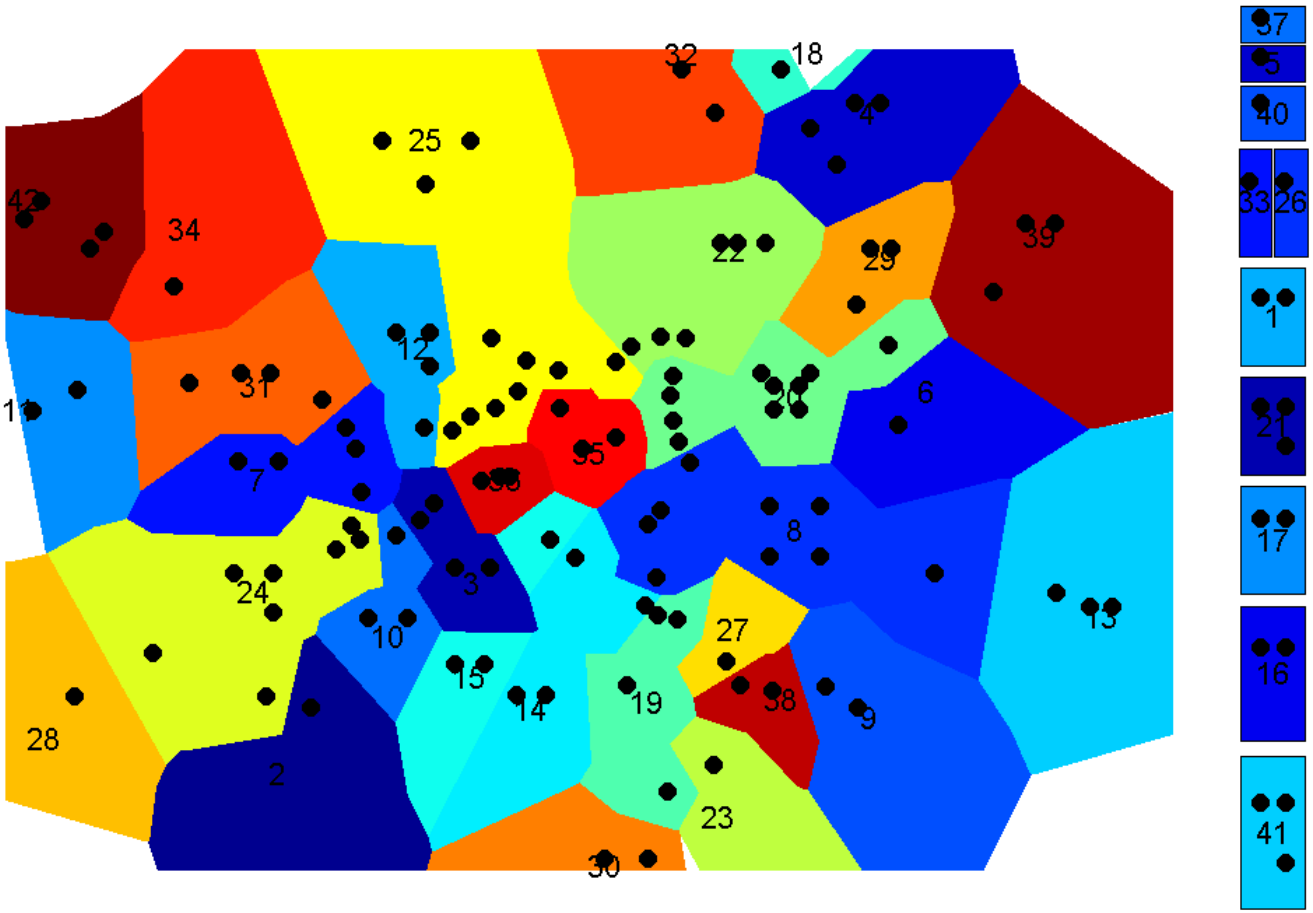
Диаграмма Вороного для множества областей знаний. Числами обозначены номера областей знаний.
Результат получен при $M = 3, L = 7.5$.



Окончательный результат

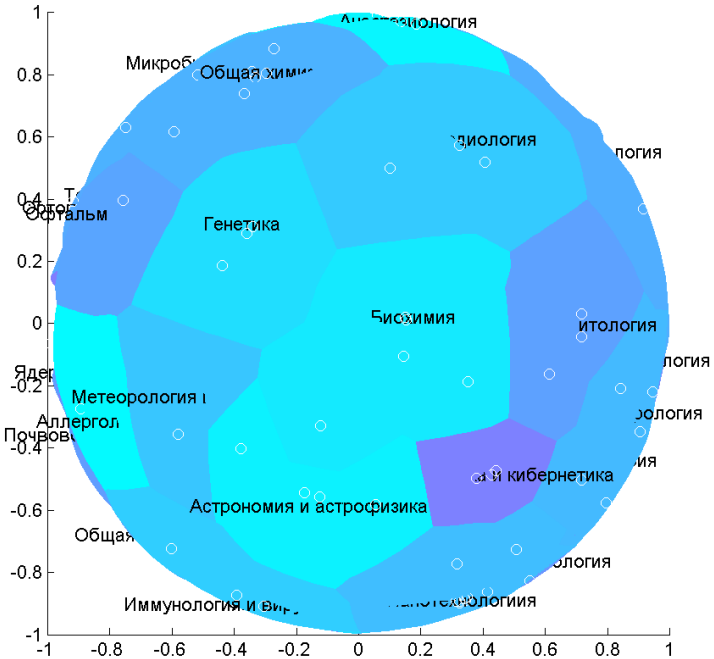
Числами обозначены номера областей знаний.

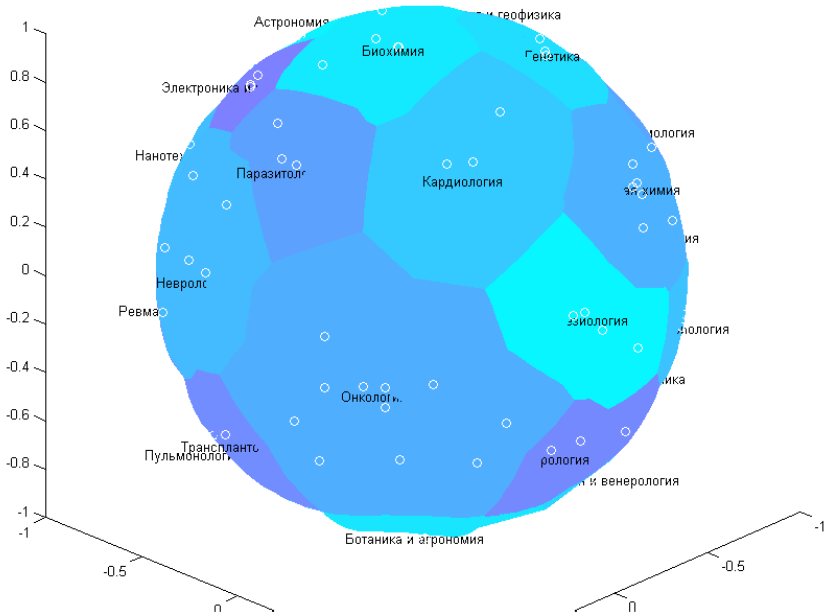
Результат получен при
$W = 1000000, V = 1000000, G = 100000, H = 1.$



Глобус для самой большой компоненты связности

Результат получен при
$M = 3, L = 7.5, W = 1000000, V = 1000000, G = 100000, H = 1.$
Ниже показан один и тот же глобус с двух разных сторон.





Заключение

- разработан новый метод визуализации данных, описываемых трехдольной полужесткой моделью;
- метод реализован;
- проведен вычислительный эксперимент на реальных данных.