Московский Физико-Технический Институт
(Государственный Университет)

Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

«Визуализация результатов картирования групп речевых маркеров»

Выполнила:
студентка 4 курса 074 группы
Вдовина Евгения Александровна

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

Москва, 2014

Аннотация

Одним из важных приемов в аналитической деятельности является визуализация исходных данных. В работе рассматривается случай реляционных данных, связывающих три единицы анализа — объекты, маркеры, классы. Соответствующая модель была названа трехдольной. Данные визуализируются в виде карты. Метод визуализации разработан согласно классической методике. Формализация задачи базируется на требованиях экспертов, решение задачи основывается на методе раскладки графа.

Содержание

Введение

Людям довольно часто приходится работать с каким-то способом классифицированными объектами. Например, это могут быть тексты, рассортированные по темам, такие как статьи в научных журналах, тезисы конференций, электронные сообщения и многое другое.

Если рубрик и объектов достаточно много, то человеку становиться трудно из таблиц или списков получить «общее впечатление» о связях между ними. Гораздо удобнее было бы каким-то образом визуализировать распределение объектов по рубрикам, и тогда можно было бы легко оценить картину в целом.

Задача визуализации данных существует не первый день, и придумано множество способов ее решения. Но данные по своей структуре бывают разными, и придумать метод, который бы отлично визуализировал любые данные, невозможно. Наиболее общие методы, приминимые ко всем структурам данных, грешат тем, что показывают только один параметр, который человек сам выбирает и считает перед визуализацией. Поэтому были созданы более специализированные методы для самых распространенных типов данных. В то же время эти методы, как правило, остаются довольно общими, т. к. опираются только на самые общие свойства того типа данных, для которого приспособлены.

Однако, если имеются данные довольно специфической структуры и нужно показать на диаграмме довольно много параметров, отражающих особенности этого типа данных, то очень тяжело подобрать метод, который бы справился с этой задачей. В данной работе рассмотрена трехдольная полужесткая модель данных на примере картирования областей знаний в наукометрии и разработан метод ее визуализации на основе существующих алгоритмов.

Глава 1

Трехдольная полужесткая модель данных

В этой главе описывается модель данных, с которой идет работа впоследствии.

1.1 Описание модели

Определение 1. *Трехдольная полужесткая модель — это модель гетерогенных реляционных данных, в которой присутствуют три единицы анализа — объекты, маркеры и классы, причем одно из соответствий между единицами анализа является функциональным, то есть представляет собой отношение «один ко многим» (функцию, отображение).*

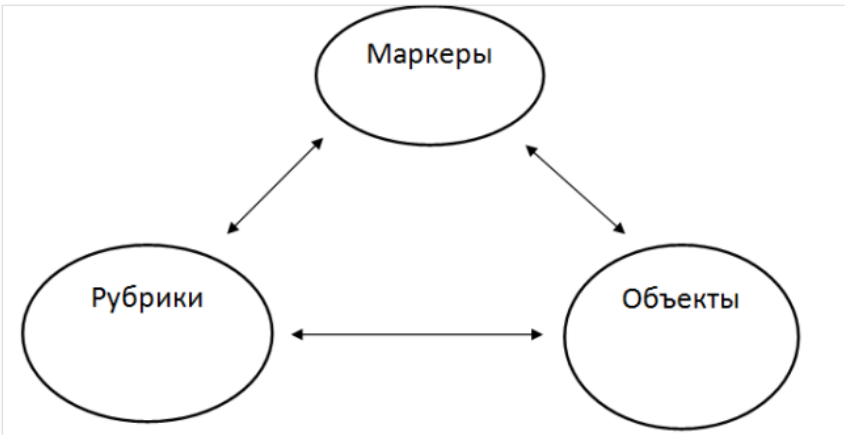


Рис. 1.1: Схема модели

Так как описываемые этой моделью данные являются реляционными, то их удобно хранить в реляционных базах данных.

Обычно эта модель используется в предположении, что одно из соответствий между единицами анализа является композицией двух других. Если данные хранятся в реляционной базе данных, это отношение не хранится. Например, если композицией

является отношение, связывающее объекты и классы, то схема модели выглядит, как на рисунке (1.2)

Рис. 1.2: Схема модели. Отношение между объектами и классами является композицией отношений между объектами и маркерами и между маркерами и классами.

Ниже приведено несколько примеров данных из разных предметных областей, которые могут быть описаны с помощью трехдольной полужесткой модели.

Пример 1. Предметная область — социология, данные — коллекция интервью. В качестве объектов выступают документы, каждый из которых соответствует одному разделу интервью одного респондента. Маркерами в данном случае являются слова в документах, а классами — разделы интервью. Набор разделов интервью у всех респондентов одинаков.

Пример 2. Научные конференции. Здесь объекты — это документы, представленные на конференцию, маркеры — ключевые термины, выделенные экспертами, классы — секции и подсекции конференции.

Пример 3. Научные публикации. В этом случае объектами являются публикации, маркерами — ключевые слова, классами — области знаний.

1.2 Типичные задачи на данной модели

В данной работе решается исключительно задача визуализации данных, описываемых трехдольной полужесткой моделью. Однако это не единственная задача, которую на таких данных можно поставить. Для лучшего понимания модели читателем ниже приводятся другие задачи на трехдольной полужесткой модели данных:

- по двум известным связям между единицами анализа найти третью;
- восстановить/обогатить связи, если не все существующие указаны или добавились новые элементы:
  - между ключевыми словами и документами (не везде указаны ключевые слова);
  - между ключевыми словами и рубриками (не все ключевые слова привязаны к рубрикам);
- извлечь из текста кандитатов в ключевые слова;

- извлечь словари для рубрик;
- построить новые связи (связать журналы с областями знаний)

В **Примере 1** про коллекцию интервью из предыдущего подраздела известны отношения между документами и разделами и между словами и документами. Нужно найти отношение между словами и разделами и выделить термины, наиболее хорошо описывающие раздел, то есть определить лексические ядра рубрик.

Пример 2. На данных о научных конференциях известны отношения между словами и документами и между документами и рубриками. Нужно распределить термины по секциям.

Пример 3. На данных о научных публикациях изветсны отношения между объектами и маркерами (многие ко многим) и между объектами и рубриками (многие к одному). Нужно маркеры приписать к рубрикам.

1.3 Данные из наукометрии

Результаты исследования будут проиллюстрированы на примере реальных данных из области наукометрии.

В этих данных следующие три единицы анализа:

- объкеты — публикации;
- маркеры — ключевые слова (речевые маркеры);
- классы — области знаний (рубрики).

Имеются следующие отношения между единицами анализа:

- объекты – маркеры (многие ко многим);
- маркеры – классы (многие к одному);
- объекты – классы (композиция двух предыдущих отношений).

Предполагается, что данные чистые:

- каждая публикация связана хотя бы с одним речевым маркером;
- каждый речевой маркер связан хотя бы с одной публикацией;
- каждый речевой маркер связан ровно с одной рубрикой;
- каждая рубрика связана хотя бы с одним речевым маркером.

В имеющихся данных 42 рубрики, 133 ключевых слова и 1756 публикаций.

Глава 2

Общая задача визуализации

Общая задача визуализации состоит в том, чтобы исходные данные представить в виде изображения. Это изображение называется диаграммой. Диаграммы в основном состоят из геометрических объектов (точек, линий, фигур различной формы и цвета) и вспомогательных элементов (осей координат, условных обозначений, заголовков и т. п.). Геометрические объекты, из которых состоит диаграмма, называются единицами отображения. По тому, какие именно единицы отображения используются в диаграмме, диаграммы делятся на различные типы. Обычно тип диаграммы входит в постановку задачи, то есть нужно представить исходные данные в виде диаграммы определенного типа.

Чтобы диаграмма была полезной, она должна отвечать некоторым требованиям. Во-первых, она должна содержать нужную информацию. Это условие достигается путем установления связи единиц отображения и их числовых характеристик с исходными данными. Во-вторых, она должна быть понятной и хорошо читаемой, а также приятной глазу. Итак, к диаграмме предъявляются два типа требований:

- смысловые, зависящие от данных;
- эстетические, от данных не зависящие.

В качестве примера можно привести стандартные диаграммы, такие как столбчатая и круговая. Они могут быть очень красивыми, но показываеют всего один-два числовых параметра: рисунки (2.1) и (2.2).

Рис. 2.1: Круговая диаграмма для данных, описанных в разделе 1.3

Рис. 2.2: Столбчатая диаграмма для данных, описанных в разделе 1.3

Глава 3

Визуализация трехдольной полужесткой модели существующими методами

Целью данной главы является демонстрация результатов применения известных методов визуализации к данной модели.

3.1 Визуализация методами раскладки графов

Реляционные данные очевидным образом представляются в виде графа. Поэтому в этом разделе для визуализации трехдольной полужесткой модели применяются методы раскладки графов.

3.1.1 Некоторые методы раскладки графов

Обычно графы визуализируют следующим образом: на плоскости вершины обозначают точками, ребра — линиями, соединяющими соответсвующие точки. Если граф небольшой или у него мало ребер, то все просто: даже не очень аккуратная картинка будет понятной. Чем больше граф и чем более он близок к полному, тем сложнее становиться создать понятную картинку.

Один из методов решения этой проблемы описан в [9]. Идея метода — использовать тот же принцип, что и в кабельных сетях: когда проводов становится слишком много, их объединяют в жгуты. В начале фиксируется расположение вершин графа, затем строятся ребра. При этом постороение ребра происходит в несколько этапов. Сначала находится путь от одной вершины ребра к другой через другие ребра. Затем через полигон, образованный вершинами, входящими в путь, проводится кривая. Авторы этого метода пришли к выводу, что лучше всего для целей визуализации в качестве кривых подходят кусочно-заданные кубические B-сплайны.

Другой метод изображает ребра прямыми линиями, и хорошее качество картинки достигается подбором оптимального положения вершин. Force-directed graph

drawing – это класс алгоритмов для визуализации графов эстетически приятным способом [12, 17, 7]. Задача расстановки вершин решается с помощью создания системы сил между множеством вершин и множеством ребер, базирующейся на их взаимном расположении. Далее система сил используется для симуляции движения вершин и ребер или же энергия полученной «физической» системы минимизируется.

При создании «физической» системы обычно силы притяжения, подобные силам упругости пружин с нулевой длиной в недеформированном состоянии, используются в сочетании с силами отталкивания, подобными кулоновским. Таким образом, узлы графа отделяются друг от друга из-за отталкивания между одноименными электрическими зарядами, а вершины, связанные друг с другом ребрами, притягиваются за счет «пружин».

Другой вариант расстановки сил заключается в том, чтобы между каждой парой вершин протянуть пружину с длиной, пропорциональной теоретически рассчитанной. Тогда не нужно отдельно вводить силы отталкивания. Далее можно минимизировать квадрат разности между «идеальной длиной» пружины и расстоянием между соответствующими вершинами. Кроме выше перечисленных сил можно использовать и другие. Также можно вводить другие законы, например силу упругости пружины можно сделать не линейной, а логарифмической.

После того, как система сил создана, можно имитировать ее поведение, как поведение физической системы. Силы будут сдвигать вершины графа, и их взаимное расположение будет меняться от итерации к итерации, пока система не достигнет положения равновесия.

Можно также непосредственно вычислить энергию системы и искать ее глобальный минимум с помощью, например, метода отжига или генетического алгоритма. Преимущества по сравнению с другими алгоритмами визуализации графов:

- Хорошие результаты

  По крайней мере для графов среднего размера (50 – 100 вершин), полученные результаты хорошо соответствовали следующим критериям: одинаковая длина ребер, равномерное распределение вершин по пространству диаграммы, видна симметрия, если она есть. Часто важно выполнение последнего критерия, но добиться его выполнения другими алгоритмами трудно.

- Гибкость

  Force-directed алгоритмы можно легко расширить и адаптировать к дополнительным эстетическим требованиям.

- Интуитивная понятность

  Так как он основан на физических аналогиях с широко известными объектами, то принципы его работы понятны и предсказать его поведение довольно легко.

- Простота

  Обычно force-directed алгоритмы просты и могут быть реализованы в несколько строчек кода.

- Интерактивность

  Визуализируя граф на разных стадиях работы алгоритма, можно получить представление о его эволюции.

Недостатки:

- Долгое время работы

  Типичный force-directed алгоритм в общем случае предполагает сложность $O(n^3)$, где n – количество вершин входного графа. Общее количество итераций оценивается в $O(n)$, но на каждой такой итерации нужно посчитать силы между всеми парами вершин.

- Проблема попадания в локальный минимум

  Часто локальный минимум бывает значительно хуже глобального. Для многих алгоритмов конечный результат сильно зависит от начального расположения вершин. Чем больше вершин у графа, тем серьезней становиться проблема локальных минимумов. Комбинация разных алгоритмов может помочь решить эту проблему. Можно разумное начальное приближение получить с помощью одного, а потом воспользоваться вторым.

3.1.2 Визуализация исходных данных

В данном разделе визуализируются без каких-либо преобразований данные, описанные в разделе 1.3. Данные представляются в виде графа следующим образом: публикациям, ключевым словам и областям знаний сопоставляются вершины, отношениям между публикациями и ключевыми словами, между ключевыми словами и областями знаний сопоставляются ребра. Плучается неориентированный невзвешенный граф, изображенный на рисунке (3.1).

3.1.3 Визуализация агрегированных данных

Исходный граф, о котором шла речь в предыдущем подразделе, можно уменьшить. Для этого нужно каким-то образом агрегировать информацию. Предлагается агрегировать информацию о публикациях. Сделать это можно следующим образом: удалить из графа вершины, соответствующие публикациям, но при этом добавить новые ребра и назначить каждому ребру вес.

В исходном графе не было ребер, соединяющих ключевое слово с ключевым словом или область знания с областью знания. Однако это не означает, что между ними не было путей. Эти пути проходили через вершины, соответствующие публикациям. Если эти вершины убрать, то пути пропадут. Чтобы этого не произошло, предлагается добавить ребра, напрямую соединяющие те ключевые слова и области знаний, между которыми были пути. Веса ребер позволят учесть количество публикаций, связанных с той или иной вершиной графа.

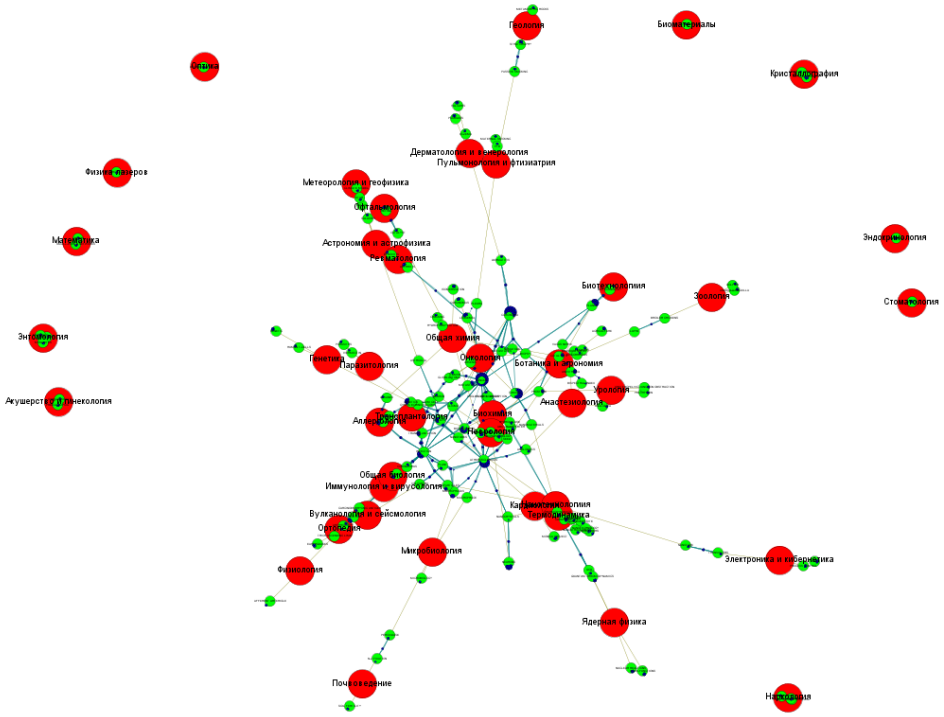


Рис. 3.1: Визуализация реальных данных методом раскладки графа. Красные, зеленые и синие вершины — области знаний, ключевые слова и публикации соответственно. Картинка получена с помощью программы gephi (алгоритм Force Atlas 2).

Таким образом, из исходного графа удаляются публикации, и добавляются ребра между ключевыми словами, имеющими общие публикации, и между областями знаний, имеющими общие публикации. Веса ребер назначаются следующим образом: ребра «ключевое слово – ключевое слово» и ребра «область знаний – область знаний» имеют вес, равный количеству общий публикаций у вершин, ребра «ключевое слово – область знаний» имеют вес, равный количеству публикаций, относящихся к соответствующему ключевому слову. Преобразованный таким образом граф показан на рисунке (3.2).

3.2 Переход к иерархической структуре

Можно заметить, что если бы оба отношения в данных, описанных в разделе 1.3, были бы отношениями «один ко многим», то граф, соответствующий этим данным (описан в подразделе 3.1.2), был бы лесом. Из леса можно получить дерево, добавив корневую вершину. В этом разделе показано, как из графа, соответствующего данным из раздела 1.3, получить дерево и как его можно визуализировать.

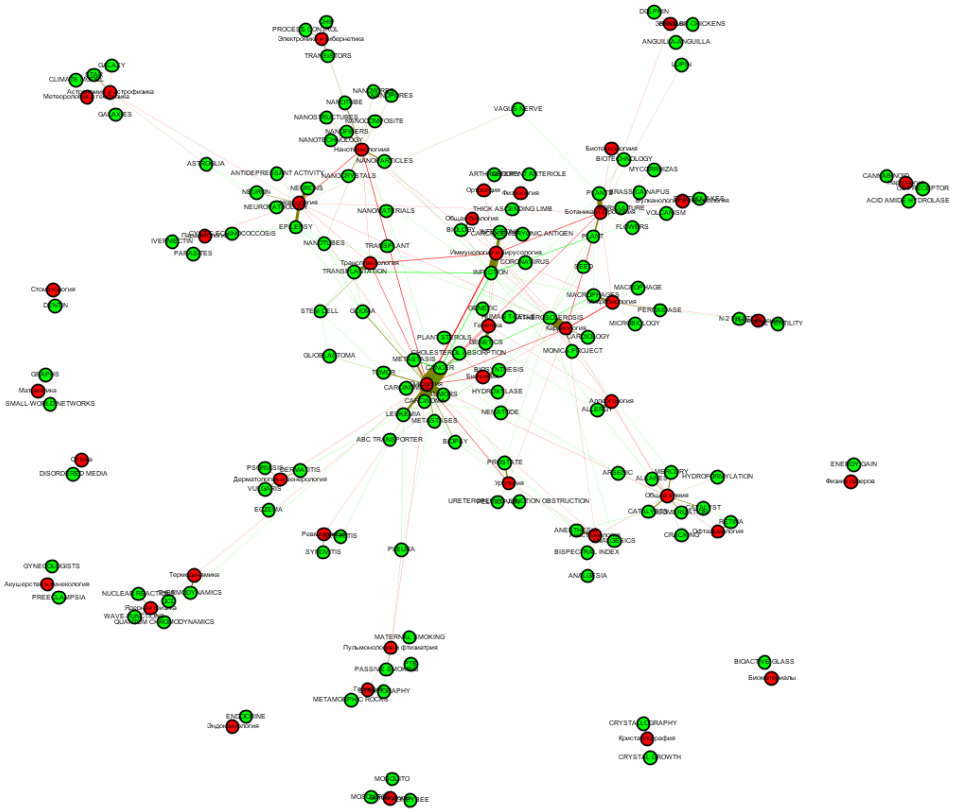


Рис. 3.2: Визуализация реальных данных методом раскладки графа. Красные и зеленые вершины — области знаний и ключевые слова соответственно. Цвет ребра получается «смешиванием» цветов его концов. Картинка получена с помощью программы gephi (алгоритм Force Atlas 2).

3.2.1 Преобразование исходного графа в дерево

В иерархической (древовидной) структуре каждый элемент относится только к одной категории на уровень выше. В трехдольной полужесткой модели это правило может нарушаться одной из единиц анализа. В данных, описанных в разделе 1.3, это правило нарушается объектами (публикациями). Если публикации убрать из рассмотрения и добавить корневую категорию, то получится структура с тремя уровнями иерархии. Убранную единицу анализа можно учесть в весах ребер полученного дерева. Нужно соответствующий вес назначить равным количеству публикаций, принадлежащих категории, соответствующей одному из концов ребра. Целесообразно выбрать тот конец ребра, которому соответствует категория более низкого уровня иерархии. Если преобразовать таким образом исходный граф, соответствующий данным, описанным в разделе 1.3, то получается дерево, изображенное на рисунке (3.3). Если в графе, соответствующем данным, несколько компонент связности, то можно добавить еще один уровень иерархии, веса ребер назначить аналогично. Тогда получится дерево, показанное на рисунке (3.4).

3.2.2 Treemap

Существует еще один метод визуализации иерархических структур, кроме визуализации их как графа-дерева. Этот метод называется treemapping.

Treemapping – это метод визуализации иерархической (древовидной) структуры данных с помощью вложенных друг в друга подобных фигур [11]. Обычно это прямоугольники. Каждая ветвь дерева представляется прямоугольником, покрытым прямоугольниками поменьше, представляющими подветви. Площадь фигур пропорциональна количеству объектов, относящихся к соответствующей ветви дерева. Цвет прямоугольников может означать принадлежность к какой-либо ветви или какое-то свойство входящих в него объектов. Этот метод не предполагает пересечений между иерархическими категориями одного уровня. Соседство прямоугольников на диаграмме ничего не обозначает. Существует несколько алгоритмов, располагающих фигуры на плоскости, некоторые из них описаны в [6]. Кроме прямоугольников можно исползовать и другие фигуры [18, 4, 16]. На данный момент описаны treemap с базовыми фигурами самой разной формы, в том числе и трехмерные [15].

3.2.3 Визуализация данных из наукометрии с помощью treemapping

В подразделе 3.2.1 исходный граф был преобразован во взвешенное дерево. При этом вес ребра, соединяющего две категории, равен количеству публикаций, относящихся к более низкой по уровню категории. Для удобства введем числовую характеристику вершины, также называемую весом. Пусть вес вершины равен количеству публикаций, относящихся к ней. Тогда в общем случае вес категории не равен сумме весов относящихся к ней категорий уровнем ниже. Это происходит, потому что одна публикация может относиться к нескольким ключевым словам и, как следствие, к нескольким областям знаний. Возникает две возможности учета публикаций:

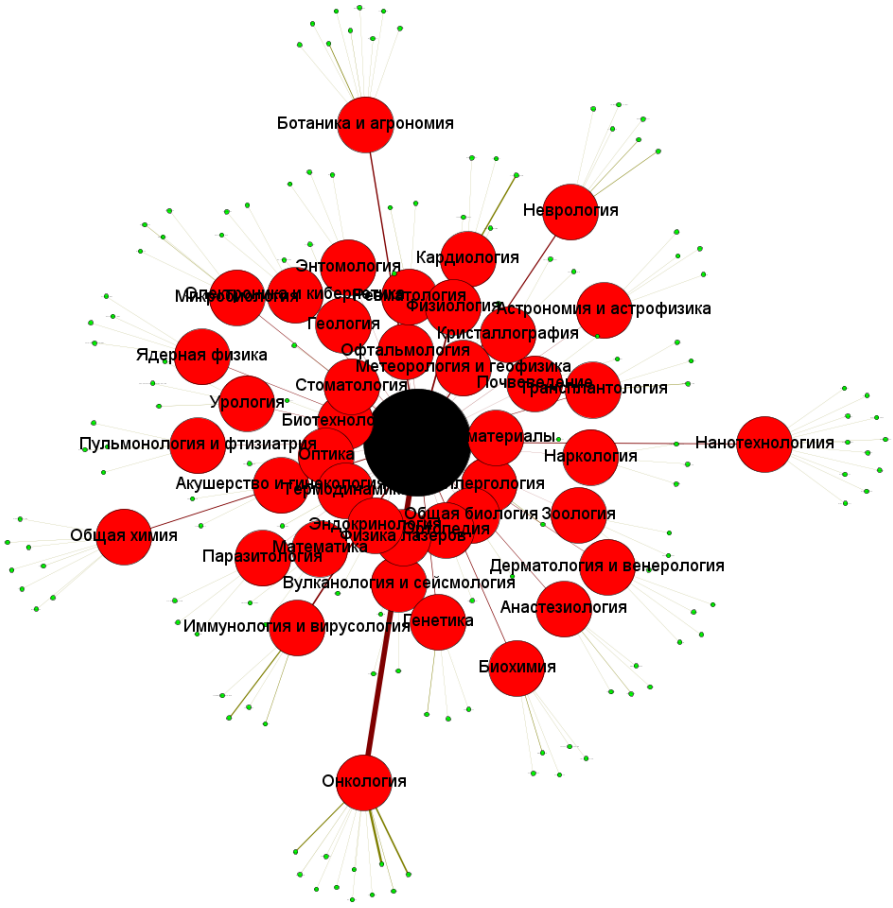


Рис. 3.3: Визуализация реальных данных методом раскладки графа. Красные и зеленые вершины — области знаний и ключевые слова соответственно. Черная вершина — корень дерева. Цвет ребра получается «смешиванием» цветов его концов. Картинка получена с помощью программы gephi (алгоритм Yifan Hu).

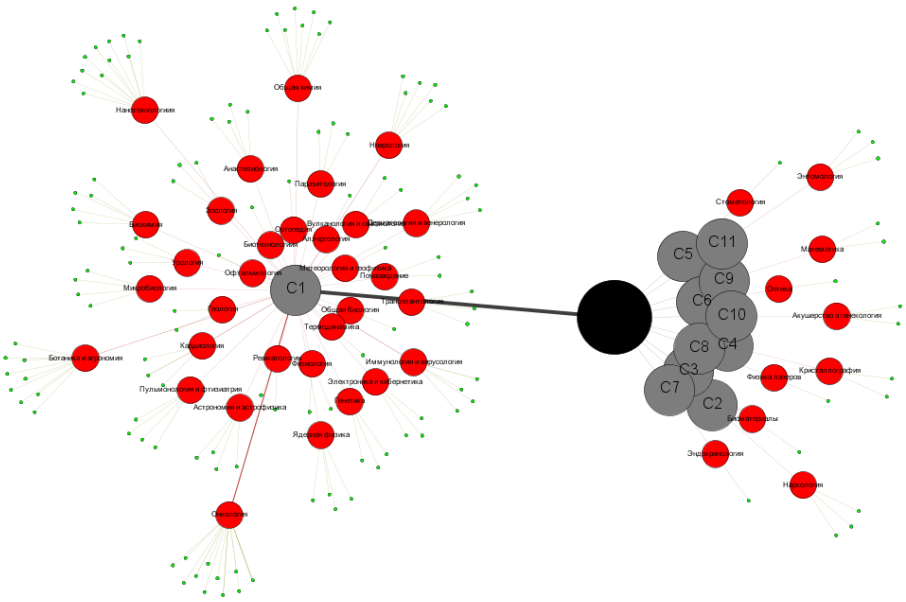


Рис. 3.4: Визуализация реальных данных методом раскладки графа. Красные, зеленые и серые вершины — области знаний, ключевые слова и компоненты связности соответственно. Черная вершина — корень дерева. Цвет ребра получается «смешиванием» цветов его концов. Картинка получена с помощью программы gephi (алгоритм Yifan Hu).

1. с повтором: если публикация относится к нескольким категориям одного уровня, то в каждой из этих категорий она учитывается с весом 1, как и любая другая публикация; тогда сумма весов категорий может превышать вес категории на уровень выше, к которой относятся эти категории;

2. с дробным весом: если публикация относится к n ключевым словам, то она учитывается в каждом ключевом слове с весом $\frac{1}{n}$; тогда вес категории равен сумме весов категорий на уровень ниже, относящихся к ней.

В данной работе используется только 1-ый способ учета публикаций. Если применить treemapping к преобразованному к дереву графу, то получаются рисунки (3.5) и (3.6).

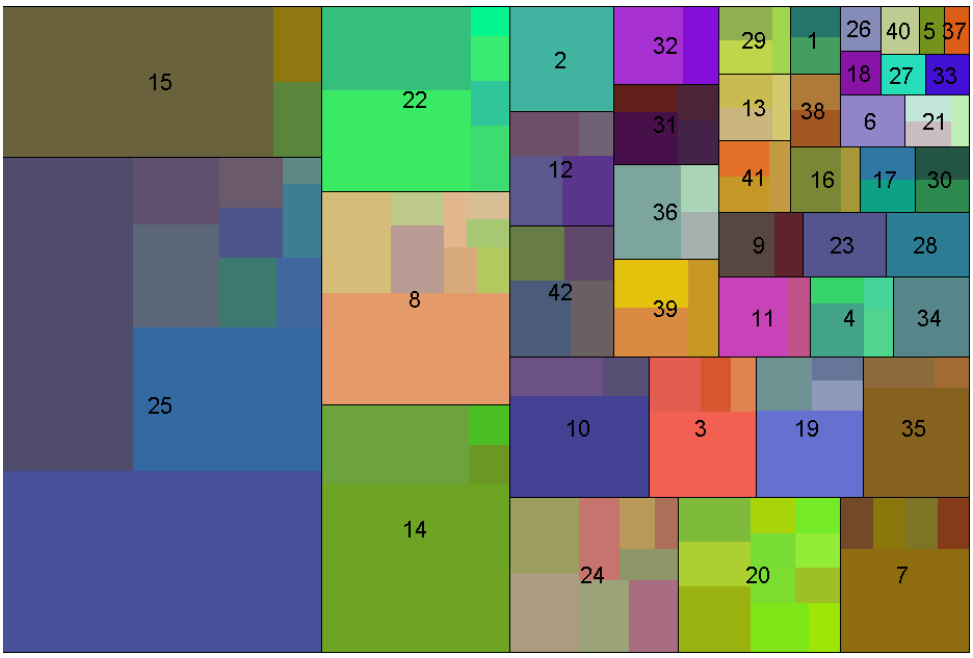


Рис. 3.5: Визуализация реальных данных с помощью treemap. Связные компоненты в графе не выделены, числа — это номера тем.

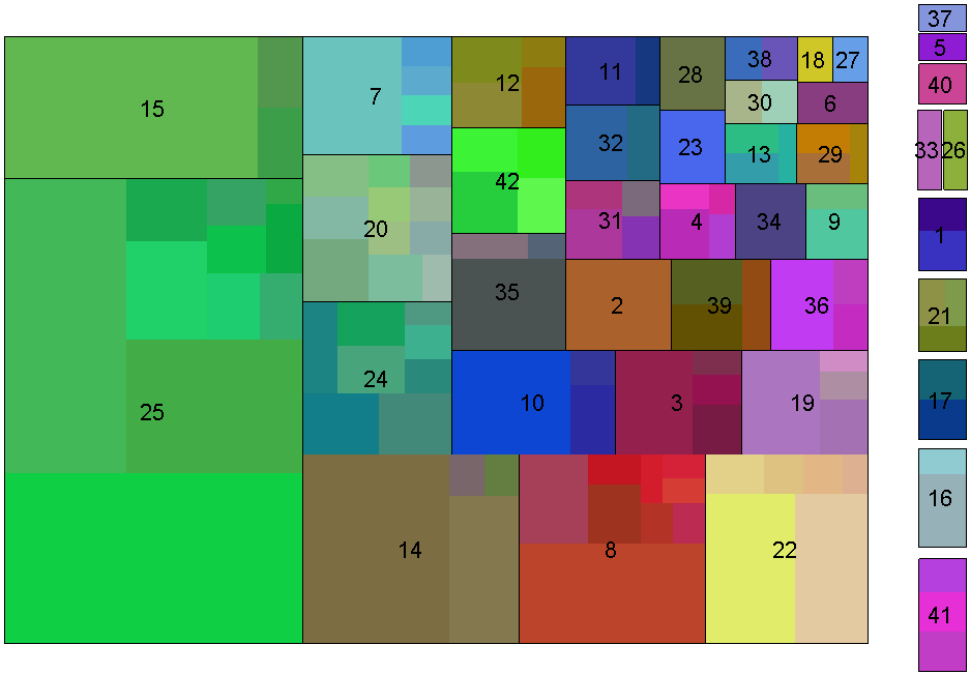


Рис. 3.6: Визуализация реальных данных с помощью treemap. Выделены связные компоненты в графе, соответствующие прямоугольники искусственно (методом treemap это не предусмотрено) уменьшены на 10% для того, чтобы между ними появились зазоры и диаграмма была более понятной. Числа — это номера тем.

Глава 4

Специальные диаграммы с площадными формами

Площадная форма — это подмножество области построения диаграммы, ограниченное замкнутой кривой. Обычно площадная форма является односвязной областью. Площадные формы на диаграмме могут пересекаться или не пересекаться. Типичным примером диаграмм с пересекающимися формами являются диаграммы Эйлера-Венна. Диаграммы с непересекающимися формами называются картами.

4.1 Диаграммы Венна

Обзор по диаграммам Эйлера-Венна можно найти в [14]. Эти диаграммы показывают отношения между конечным набором множеств. Диаграмма Венна состоит из набора замкнутых кривых, изображенных на плоскости. Обычно замкнутые кривые – это круги. Внутренность круга представляет элементы множества. Внешность круга, напротив, представляет те элементы, которые не принадлежат этому множеству. Диаграмма Венна для набора из $2-3$ множеств может быть нарисована с помощью кругов, но в плоском случае при увеличении количества множеств неизбежно некоторая потеря в симметричности фигур, представляющих множества. Диаграмма Эйлера похожа на диаграмму Венна, но она не обязана содержать все гипотетически возможные 2^n зоны (n -- количество множеств) представляющие все возможные комбинации включения и исключения для каждой компоненты множеств. Диаграмма Эйлера содержит только те зоны, которые соответствуют реально возможным отношениям между множествами в данном контексте. Пустые пересечения на диаграмме Эйлера не показаны. То есть на практике при визуализации каких-либо данных получаются именно диаграммы Эйлера, как в [3]. Обычно диаграммы Эйлера-Венна не отражают абсолютных или относительных размеров множеств, они схематичны и вся информация содержится в относительном положении множеств по отношению друг к другу. Однако, было опубликованы статьи, в том числе [19, 5], о том, как этот недостаток можно исправить.

4.2 Карта как тип диаграммы

Карта — это диаграмма, на которой могут быть следующие единицы отображения:

- непересекающиеся области (площадные формы), на политической карте мира это страны;
- точечные маркеры (могут обозначать города, автобусные остановки или другие точечные в масштабах карты объекты);
- линейные элементы (пути), например дороги и реки.

У единиц отображения есть характеристики:

- у форм — площадь, периметр, соотношение сторон, длина общих границ с другими формами, положение на карте и др.;
- у точечных маркеров — положение на карте, расстояния до других единиц отображения, размер;
- у путей — положение на карте, длина.

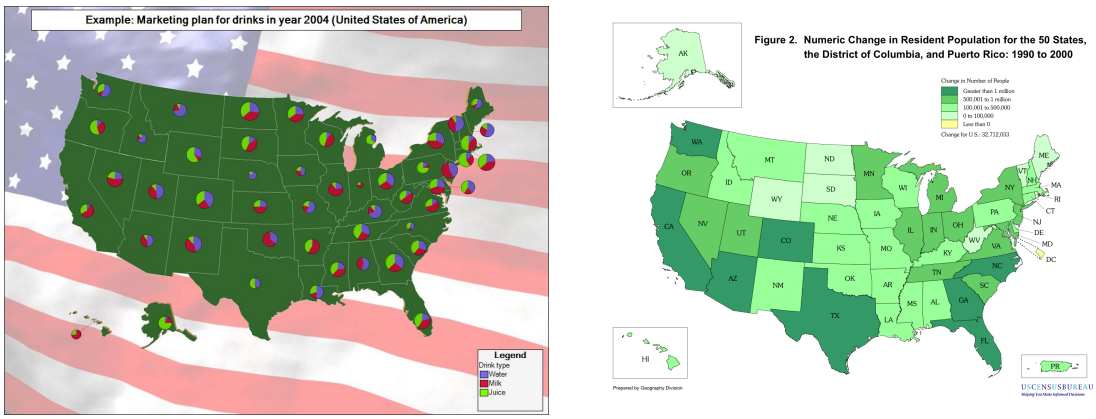
В это описание не вписывается физическая карта: на ней высота над уровнем моря указана в каждой точке, а не для целых единиц отображения.

Числовые характеристики единиц отображения могут быть экстенсивными (аддитивными) и интенсивными. К первым относятся такие параметры, как площадь, численность населения и другие. Ко вторым относится, например, плотность населения.

Не всегда на картах площадные формы заданы изначально. Иногда их нужно найти.

4.3 Карты с предопределенными областями. Картограммы(choropleth)

Примером карт, в которых формы заданы изначально, служат картограммы. Картограммы — это карты с предопределенными площадными формами, на которых показаны дополнительные числовые характеристики единиц анализа. Примером таких карт служат специальные географические карты. В них пложение, форма, площадь стран заданы извне. Дополнительная информация отображается на них с помощью цвета, градиетной окраски, небольших диаграмм поверх единиц отображения, маркеров (условных обозначений) на площадных формах [2]. Примеры катрограмм показаны на рисунках (4.1(a)) и (4.1(b)).



(a) (b)

Рис. 4.1: Примеры карт с предопределенными областями.

4.4 Карты с искажением исходных областей (cartogram)

На таких картах дополнительная информация отражается за счет изменения площади областей. Форма области приближает исходную или не зависит от неё. При этом положение области становится приблизительным [8, 10]. Пример такой карты показан на рисунке (4.2).

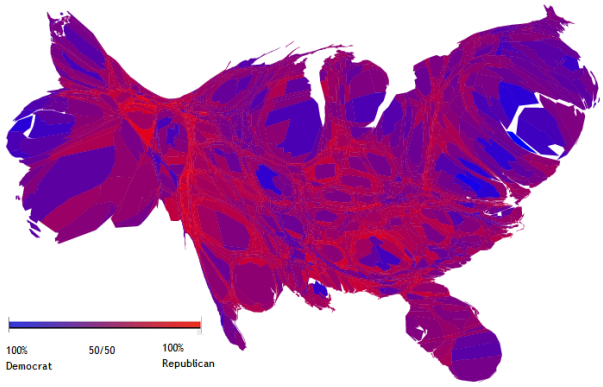


Рис. 4.2: Created by Michael Gastner, Cosma Shalizi, and Mark Newman of the University of Michigan.

Глава 5

Разработка нового метода

5.1 Методика разработки нового метода

Существует классическая методика разработки нового метода визуализации данных. Она включает следующие этапы:

1. Описание входных данных.
2. Сбор требований к диаграмме у экспертов — предполагаемых пользователей метода.
3. Математическая формализация списка требований.
4. Решение математической задачи.
5. Реализация.

Данное исследование проведено в соответствии с этой методикой.

5.2 Неформальная постановка задачи

5.2.1 Описаниие входных данных

Входные данные соответствуют трехдольной полужесткой модели, и более подробно описаны в разделе 1.3.

5.2.2 Требования к диаграмме

По итогам опроса экспертов качестве типа диаграммы была выбрана карта. Областью построения карты является прямоугольник. Была выбрана следующая связь единиц анализа с единицами отображения карты:

- документам ничего не соответствует;
- каждой рубрике соответствует одна форма;

- каждому речевому маркеру соответствует один точечный маркер.

Кроме того, были собраны различные эстетические и смысловые требования к карте: некоторые из них требуется выполнить точно, другим она должна удовлетворять наилучшим образом. Эстетические требования:

- точные
  - на диаграмме могут быть пустоты — области, не входящие ни в одну форму;
  - формы являются односвязными;
- оптимизируемые
  - пропорции формы близки к 1:1.

Смысловые требования:

- точные
  - точечный маркер находится в той форме, к какой области знаний привязано соответствующее ключевое слово.
- оптимизируемые
  - площадь формы пропорциональна количеству публикаций, относящихся к соответствующей рубрике;
  - если у двух рубрик есть общие публикации, то формы граничат, и чем больше у них общих публикаций, тем длиннее их общая граница;
  - речевой маркер, связанный хотя бы с одной публикацией, относящейся не только к рубрике этого речевого маркера, должен быть «пограничным», т. е. находится возле границы своей формы.

5.3 Подход к построению карты

Как упоминалось ранее, входные данные можно представить в виде графа. В разделе 3.2 был приведен один из способов получить карту из графа — это treemapping. Однако использование treemapping не позволит выполнить большинство требований, предъявленных к карте в неформальной постановке задачи. Поэтому нужно использовать другой метод, например можно построить диаграмму Вороного (раздел 5.5 в [1]).

Определение 2. *Диаграмма Вороного конечного множества точек M на плоскости представляет собой такое разбиение плоскости, при котором каждая область этого разбиения (ячейка) образована множеством точек, более близких к одному из элементов множества M (центру ячейки, вершине), чем к любому другому элементу множества M.*

Тогда в качестве множества точек M нужно использовать множество вершин графа.

Публикациям на карте не соответствует никакая единица отображения. Следовательно, публикаций не должно быть в графе, множество вершин которого используется для построения диаграммы Вороного. Также можно заметить, что в требованиях к карте ключевые слова делятся на «пограничные» и остальные, «внутренние». Причем единственным требованием к последним является то, что они должны находиться в соответствующей площадной форме. То есть их расположение не влияет на положение других вершин графа, а наоборот, подстраивается под него. «Внутренние» ключевые слова можно расположить на карте в самый последний момент. По этой причине далее они учитываться не будут. Значит, в графе, используемом для построения диаграммы Вороного, есть только рубрики и «пограничные» ключевые слова.

Некоторые требования, предъявленные к карте, являются оптимизируемыми. Значит, для того, чтобы их выполнить, нужно использовать оптимизационный процесс. Так как требования предъявлены к карте, то их непосредственное выполнение потребует строить карту заново на каждой итерации оптимизационного процесса. Чтобы этого избежать, предлагается строить карту в следующие 3 этапа:

1. В графе, соответствующем исходным данным, информация о публикациях агрегируется так же, как в разделе 3.1.3. Полученный граф делится на компоненты связности. С помощью treemapping область диаграммы делится на подобласти — прямоугольники, соответствующие связным компонентам. На диаграмме компоненты связности отделяются друг от друга морями (соответствующий прямоугольник уменьшается на 10%, но его центр при этом остается неподвижным), то есть одна компонента связности является на карте «континентом». В последующих двух этапах каждая компонента связности рассматривается отдельно. Таким образом, задача распадается на несколько подзадач, в каждой из которых входной граф связен.

2. От каждой компоненты связности отделяются «внутренние» ключевые слова. Для полученного графа решается задача расстановки вершин.

3. Строится диаграмма Вороного, в качестве множества точек используются вершины графа из предыдущего этапа. Из ячеек диаграммы Вороного собираются площадные формы. Одна форма является объединением ячейки, соответствующей рубрике, и ячеек, соответствующих ее «пограничным» ключевым словам. Центры ячеек, соответствующих «пограничным» ключевым словам, отмечаются круглым маркером.

Такой подход сводит задачу построения карты к задаче расстановки вершин графа.

5.4 Формализация задачи расстановки вершин графа

Задача построения карты свелась к нахождению координат вершин графа. Однако имеющиеся на данный момент требования предъявляются к карте, а не к положению вершин. Значит, нужно сформулировать требования к координатам вершин, опираясь на требования к карте.

Первый и третий этапы построения карты обеспечивают выполнение некоторых требований. Ниже приведены списки оставшихся требований.

Эстетические требования:

- точные
  - формы являются односвязными;
- оптимизируемые
  - пропорции формы близки к 1:1.

Оптимизируемые смысловые требования:

- площадь формы пропорциональна количеству публикаций, относящихся к соответствующей рубрике;
- если у двух рубрик есть общие публикации, то формы граничат, и чем больше у них общих публикаций, тем длиннее их общая граница;
- речевой маркер, связанный хотя бы с одной публикацией, относящейся не только к рубрике этого речевого маркера, должен быть «пограничным», т. е. находится возле границы своей формы.

Так как есть эстетическое требование, согласно которому формы должны быть близки к кругам, то можно задачу расположения форм приближенно рассмотреть как задачу расположения кругов на плоскости. Тогда каждой рубрике можно приписать числовую характеристику, условно называемую радиусом. С помощью этого радиуса и формулы площади круга оценивается площадь формы: $s_i = \pi R_i^2$, R_i — радиус i-ой рубрики. Площадь формы пропорциональна количеству публикаций, к ней относящихся:$s_i = \alpha^2 d_i$. Из этого следует, что $R_i = \alpha\sqrt{d_i}$, где R_i — радиус i-ой рубрики, d_i — количество публикаций, относящихся к i-ой рубрике, α — коэффициент пропорциональности. Сумма площадей всех форм на подобласти диаграммы должна равняться площади подобласти диаграммы S, отсюда:

$$\sum_i s_i = \sum_i \pi R_i^2 = \sum_i \pi \alpha^2 d_i = S \Rightarrow \alpha = \sqrt{\frac{S}{\pi \sum_i d_i}}.$$

Таким образом, α — нормировочный коэффициент, отвечающий за то, чтобы все формы поместились на диаграмме. Аналогичную числовую характеристику — радиус — можно приписать и ключевому слову.

Такой подход к расположению форм помогает выполнить требования к их площади, но совершенно бесполезен для определения длины их общих границ. Более точно определить границу формы позволяет расстановка ее «пограничных» ключевых слов. Тогда «пограничные» ключевые слова нужно располагать так, чтобы они растягивали общую границу между формами в зависимости от того, сколько у соответствующих рубрик общих публикаций.

Исходя из этих соображений, можно сформулировать следующие требования к положению вершин графа:

1. Две рубрики не могут находиться на расстоянии меньшем, чем сумма их радиусов. Если формы граничат, то расстояние между ними в точности равно сумме радиусов соответствующих рубрик. Расстояние между рубриками, не имеющими общих публикаций (то есть соответствующие им формы не должны граничить), больше суммы их радиусов.

2. Так как каждый маркер находится на соответствующей его рубрике площадной форме, то расстояние от рубрики до своего маркера должно быть не больше радиуса формы, а от рубрики до чужого маркера — больше радиуса.

3. Пусть P_i — множество публикаций, связанных с i-ым ключевым словом, а FP_i — множество публикаций, связанных с i-ым ключевым словом и относящихся одновременно к рубрикам, не связанным с i-ым ключевым словом. Аналогично радиусу рубрики, радиус ключевого слова $r_i = \alpha\sqrt{q_i}$, $q_i = |FP_i|$. Тогда, если $g_{ij} = |P_i \bigcap P_j| > 0$, то i-ое и j-ое ключевые слова должны находится около одной и той же границы на расстоянии друг от друга, равном сумме их радиусов. Таким образом «пограничные» ключевые слова будут «растягивать» границы там, где это нужно. Если $g_{ij} = |P_i \bigcap P_j| = 0$, то ключевые слова находятся на расстоянии большем, чем сумма их радиусов.

Требование 1 и 2 в совокупности гарантируют односвязность каждой формы.

Непосредственная запись требования 1 в виде оптимизационной задачи дает

$$\sum_{i,j} \chi(\mathbf{y}_i, \mathbf{y}_j) \to \min, \tag{5.1}$$

где

$$\chi(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} M\left(\rho(\mathbf{y}_i, \mathbf{y}_j) - (R_i + R_j)\right)^2, \text{ формы } i \text{ и } j \text{ граничат;} \\ L\left(\rho(\mathbf{y}_i, \mathbf{y}_j) - (R_i + R_j)\right)^2, \text{ формы } i \text{ и } j \text{ не граничат и } \rho(\mathbf{y}_i, \mathbf{y}_j) < C(R_i + R_j); \\ 0, \text{ формы } i \text{ и } j \text{ не граничат и } \rho(\mathbf{y}_i, \mathbf{y}_j) \geqslant C(R_i + R_j). \end{cases} \tag{5.2}$$

$C \geqslant 1, L, M$ — некоторые константы, $\rho(\mathbf{y}, \mathbf{x})$ — евклидова метрика, R_i — радиус i-ой рубрики. Чтобы формализовать остальные требования, используется force-directed подход. Согласно этому подходу нужно создать такую «физическую» систему, которая бы вела себя нужным образом, и минимизировать ее потенциальную энергию. В данном случае в системе должны быть следующие взаимодействия:

- рубрика отталкивает чужие речевые маркеры, если расстояние до них меньше порогового (требование 2);
- рубрика притягивает свои речевые маркеры, если расстояние до них больше порогового (требование 2);
- рубрики отталкиваются друг от друга (требование 1);
- речевые маркеры, имеющие общие публикации, притягиваются (требование 3);
- все речевые маркеры отталкиваются друг от друга (требование 3).

Потенциальная энергия взаимодействий в «физической» системе:

- Питяжение речевого маркера к «своей» рубрике (требование 2):

$$f(\mathbf{y}_k, \mathbf{x}_i) = \begin{cases} \left(\frac{\rho(\mathbf{y}_k, \mathbf{x}_i)}{R_k}\right)^8 - 1, \text{речевой маркер } i \text{ относится к рубрике } k \text{ и} \rho(\mathbf{y}_k, \mathbf{x}_i) > R_k \\ 0, \text{речевой маркер } i \text{ относится к рубрике } k \text{ и} \rho(\mathbf{y}_k, \mathbf{x}_i) \leqslant R_k \end{cases} \tag{5.3}$$

- Отталкивание речевого маркера от «чужих» рубрик (требование 2):

$$f(\mathbf{y}_k, \mathbf{x}_i) = \begin{cases} -\ln \frac{\rho(\mathbf{y}_k, \mathbf{x}_i)}{R_k}, \text{речевой маркер } i \text{ не относится к рубрике } k \text{ и} \rho(\mathbf{y}_k, \mathbf{x}_i) < R_k \\ 0, \text{речевой маркер } i \text{ не относится к рубрике } k \text{ и} \rho(\mathbf{y}_k, \mathbf{x}_i) \geqslant R_k \end{cases} \tag{5.4}$$

- Притяжение речевых маркеров, имеющих общие публикации (требование 3): $g_{ij}\rho(\mathbf{x}_i, \mathbf{x}_j)^2$
- Отталкивание маркеров друг от друга (требование 3): $\frac{r_i + r_j}{\alpha\rho(\mathbf{x}_i, \mathbf{x}_j)}$.
- Отталкивание между рубриками (требование 1): $\frac{R_i + R_j}{\alpha\rho(\mathbf{y}_i, \mathbf{y}_j)}$.

g_{ij} — число публикаций, связанные одновременно с речевыми маркерами i и j;
r_i — радиус i-ого речевого маркера, R_i — радиус i-ой рубрики, α — нормировочный коэффициент.

В итоге получился следующий функционал:

$$W\sum_{k,i} f(\mathbf{y}_k, \mathbf{x}_i) + \sum_{i,j}\left(Hg_{ij}\rho(\mathbf{x}_i, \mathbf{x}_j)^2 + G\frac{r_i + r_j}{\alpha\rho(\mathbf{x}_i, \mathbf{x}_j)}\right) + V\sum_{i,j}\frac{R_i + R_j}{\alpha\rho(\mathbf{y}_i, \mathbf{y}_j)} \to \min, \tag{5.5}$$

где W, H, G, V — некоторые числовые коэффициенты;
$\mathbf{x}_i, \mathbf{y}_k$ — векторы координат i-ого речевого маркера и k-ой рубрики соответственно;
$R_k = \alpha\sqrt{d_k}$, $r_i = \alpha\sqrt{q_i}$, $\alpha = \sqrt{\frac{S}{\pi\sum_{k=1} d_k}}$, S — площадь подобласти диаграммы;
g_{ij} — число публикаций, к которым одновременно относятся ключевые слова i и j;
d_k — количество публикаций, относящихся к k-ой рубрике;

q_i — количество публикаций у речевого маркера i, относящихся не только к той рубрике, к которой он относится.

Если в функционале (5.5) оставить только слагаемые, влияющие на положение ключевых слов, получится следующая задача:

$$W \sum_{k,i} f(\mathbf{y}_k, \mathbf{x}_i) + \sum_{i,j} \left(H g_{ij} \rho(\mathbf{x}_i, \mathbf{x}_j)^2 + G \frac{r_i + r_j}{\alpha \rho(\mathbf{x}_i, \mathbf{x}_j)} \right) \to \min . \tag{5.6}$$

Глава 6

Вычислительный эксперимент

Вычислительный эксперимент проводится на данных, описанных в главе 1.3. Задачу оптимизации имеет смысл решать только тогда, когда рубрик больше одной. В используемых данных только в одной компоненте связности больше одной рубрики. Было опробовано несколько подходов к решению оптимизационной задачи:

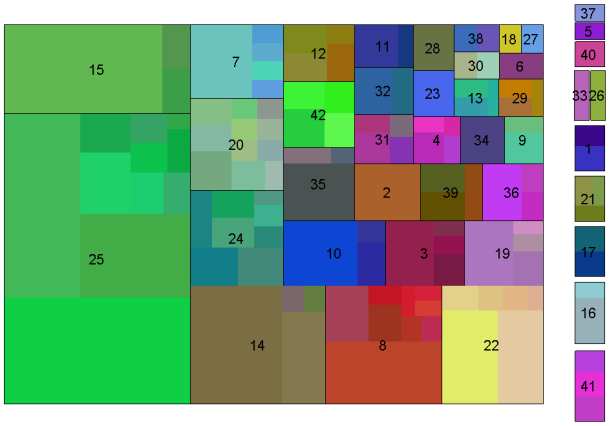


Рис. 6.1: Этап разделения очищенного графа на компоненты связности дает следующую картину (чтобы сделать картинку более понятной, прямоугольники, соответствующие связным компонентам, были уменьшены на 10% и между ними появились зазоры)

1. В качестве начального приближения для координат ключевых слов и рубрик берутся координаты центров соответствующих прямоугольников, построенных с помощью treemapping. Если координаты рубрики совпадают с координатами ключевого слова, то ключевое слово сдвигается. Далее решается оптимизационная задача (5.5), в процессе решения которой изменяются координаты и рубрик, и ключевых слов.

2. В качестве начального приближения для координат рубрик берутся координаты центров соответствующих прямоугольников, построенных с помощью treemapping. Далее решается оптимизационная задача (5.1), в процессе решения которой изменяются только координаты рубрик. Затем для каждой рубрики

определяется наименьшее расстояние до другой рубрики, оно делится пополам и внутри круга с таким радиусом располагаются пограничные ключевые слова этой рубрики. Полученные таким образом координаты ключевых слов используются как начальное приближение к задаче (5.6), где изменяются только координаты ключевых слов, а координаты рубрик остаются постоянными.

3. В качестве начального приближения для координат рубрик берутся координаты центров соответствующих прямоугольников, построенных с помощью treemapping. Далее решается оптимизационная задача (5.1), в процессе решения которой изменяются только координаты рубрик. Затем для каждой рубрики определяется наименьшее расстояние до другой рубрики, оно делится пополам и внутри круга с таким радиусом располагаются пограничные ключевые слова этой рубрики. Все полученные координаты используются как начальное приближение для решения задачи (5.5), где меняются координаты и рубрик, и ключевых слов.

Следует ввести некоторые обозначения, которые будут использоваться далее.

$$SE = \sum_{i \neq j} \phi(\mathbf{y}_i, \mathbf{y}_j), \text{ где } \phi(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} \left(\rho(\mathbf{y}_i - \mathbf{y}_j) - (R_i + R_j)\right)^2, \text{ формы } i \text{ и } j \text{ граничат;} \\ 0, \text{ иначе} \end{cases},$$

— вводится для сравнения качества расположения рубрик относительно друг друга при различных значениях числовых параметров. E, TE, KE — значение функционала (5.5), (5.1) или (5.6) соответственно, на котором произошла остановка решения оптимизационной задачи.

Сначала был применен 1-ый подход. После нескольких экспериментов стало ясно, что есть следующие проблемы: ключевые слова "отрываются"от рубрик и формы получаются несвязные, кроме того, ключевые слова стремятся слишком сильно сблизиться. С помощью изменения числовых коэффициэнтов (W, H, G, V), с которыми складываются части функционала, отвечающие за разные требования, были увеличены отталкивание между ключевыми словами, взаимодействие ключевых слов с рубриками и отталкивание между рубриками, но полностью исправить ситуацию не удалось.

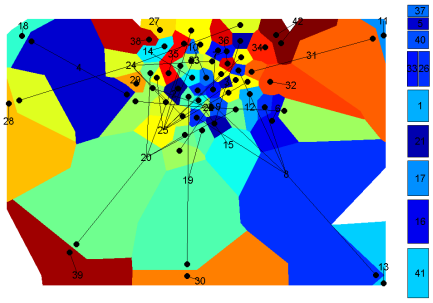
2-ой и 3-ий подходы используют расстановку рубрик без учета ключевых слов, так что можно сначала подобрать параметры в функционале (5.1). Для сравнения качества расстановки рубрик при различных числовых параметрах используется еще одна величина:

$$N = \sum_{i<j} [\text{формы } i \text{ и } j \text{ должны граничить, но не граничат}].$$

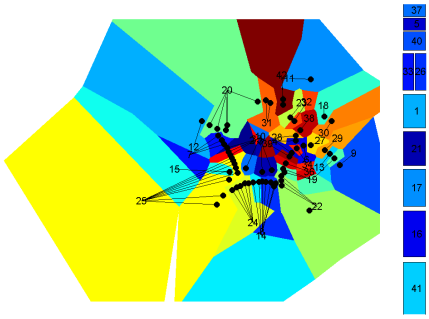
При применении 2-ого и 3-его подходов проблемы возникают все те же: ключевые слова отделяются от рубрик. Соответственно, методы борьбы с ними также остаются те же: увеличить взаимодействие между речевыми маркерами и рубриками, отталкивание между маркерами и между рубриками. Тем не менее, самым успешным оказался 3-ий подход. Ниже приведены некоторые результаты экспериментов

с указанием, с помощью какого подхода и при каких значениях числовых параметров они получены. Если на рисунке точечный маркер соединен с номером рубрики линией, то соответствующее ключевое слово относится к этой рубрике. Эти линии нужны, чтобы было понятно, связные формы получились или нет. На всех картинках, кроме рисунка (6.2), точечные маркеры обозначают «пограничные» ключевые слова, а «внутренние» ключевые слова не показаны.

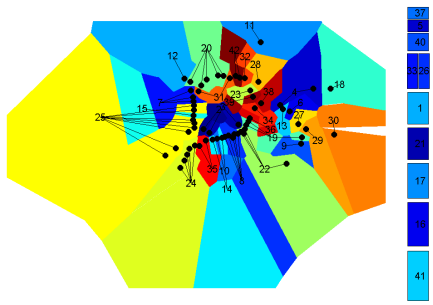
На рисунке (6.2) приведен лучший из полученных результатов. На нем показаны как «пограничные», так и «внутренние» ключевые слова. Вспомогательных линий, соединяющих «пограничные» ключевые слова с их рубриками, на рисунке (6.2).



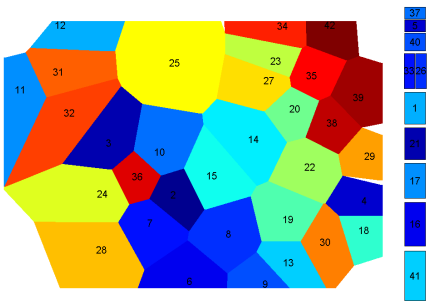
(a) Конечный результат. Использовался 1-ый подход к решению задачи оптимизации. $W = 1, H = 1, G = 100000, V = 1, SE = 1.2056 \times 10^6, E = 8.5544 \times 10^6$



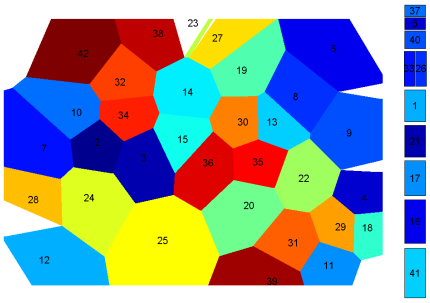
(b) Конечный результат. Использовался 1-ый подход к решению задачи оптимизации. $W = 1000000, H = 1, G = 100000, V = 1, SE = 8.5724 \times 10^5, E = 1.9534 \times 10^7$



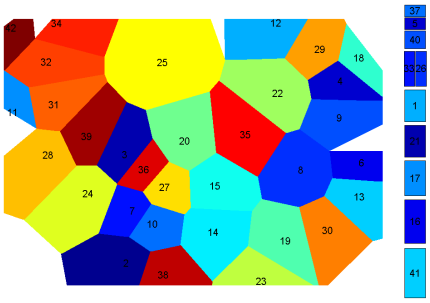
(c) Конечный результат. Использовался 1-ый подход к решению задачи оптимизации. $W = 1000000, H = 1, G = 100000, V = 100000, SE = 9.3538 \times 10^5, E = 2.1649 \times 10^7$



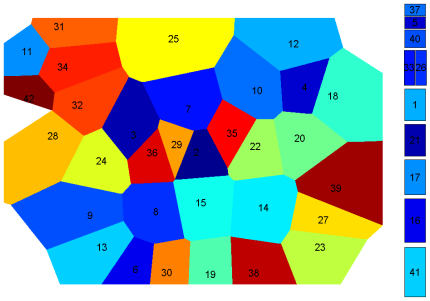
(d) Расстановка рубрик без учета ключевых слов. $M = 1, L = 1, C = 1, SE = 2.0995 \times 10^5, TE = 2.7921 \times 10^5, N = 24$



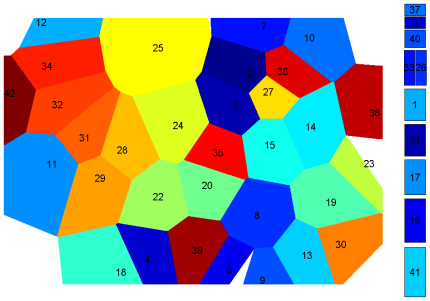
(e) Расстановка рубрик без учета ключевых слов. $M = 1, L = 7.5, C = 1, SE = 4.0169 \times 10^5, TE = 4.2918 \times 10^5, N = 23$



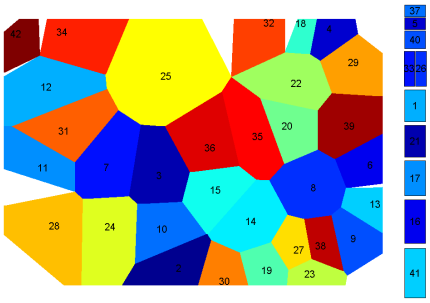
(f) Расстановка рубрик без учета ключевых слов. $M = 4, L = 7.5, C = 1, SE = 2.4856 \times 10^5, TE = 1.1889 \times 10^6, N = 18$



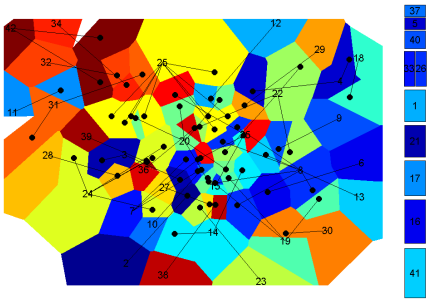
(g) Расстановка рубрик без учета ключевых слов. $M = 4, L = 7.5, C = 1.1, SE = 2.8409 \times 10^5, TE = 1.5199 \times 10^6, N = 23$



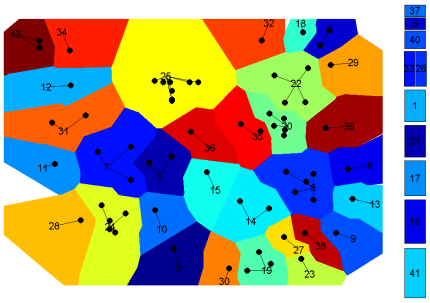
(h) Расстановка рубрик без учета ключевых слов. $M = 5, L = 7.5, C = 1, SE = 2.2972 \times 10^5, TE = 1.4985 \times 10^6, N = 23$



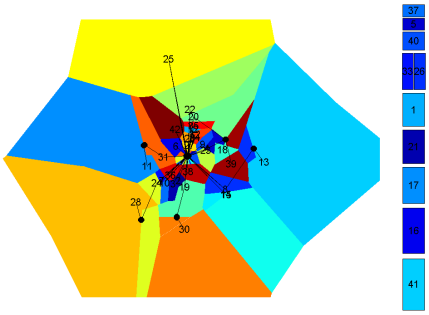
(i) Расстановка рубрик без учета ключевых слов. $M = 3, L = 7.5, C = 1, SE = 2.3317 \times 10^5, TE = 8.3657 \times 10^5, N = 19$



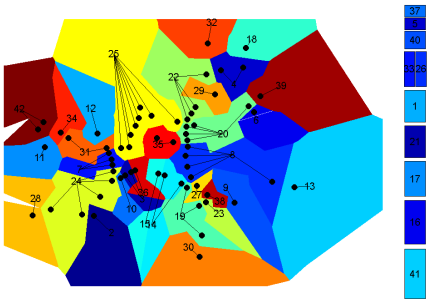
(j) Конечный результат. Использовался 2-ой подход к решению задачи оптимизации. $M = 4, L = 7.5, C = 1, W = 1, H = 1, G = 1, KE = 3.5764 \times 10^7$



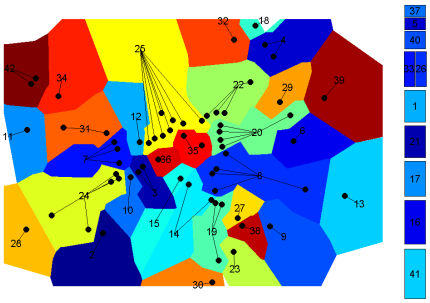
(k) Конечный результат. Использовался 2-ой подход к решению задачи оптимизации. $M = 3, L = 7.5, C = 1, W = 10000, H = 1, G = 1, KE = 3.8229 \times 10^7$



(l) Конечный результат. Использовался 3-ий подход к решению задачи оптимизации. $M = 3, L = 7.5, C = 1, W = 10000, H = 1, G = 1, V = 1, SE = 1.0144 \times 10^6, E = 1.0200 \times 10^4$



(m) Конечный результат. Использовался 3-ий подход к решению задачи оптимизации. $M = 3, L = 7.5, C = 1, W = 1000000, H = 1, G = 100000, V = 100000, SE = 5.7539 \times 10^5, E = 1.5661 \times 10^7$



(n) Конечный результат. Использовался 3-ий подход к решению задачи оптимизации. $M = 3, L = 7.5, C = 1, W = 1000000, H = 1, G = 100000, V = 1000000, SE = 4.6256 \times 10^5, E = 3.2935 \times 10^7$

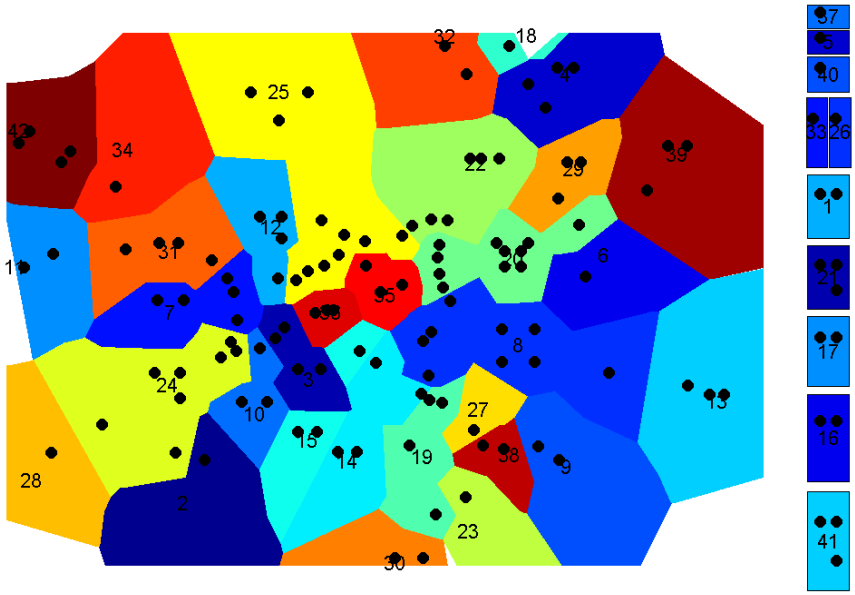


Рис. 6.2: Лучший результат. Получен с помощью 3-его подхода к решению задачи задачи оптимизации при следущих числовых параметрах: $M = 3, L = 7.5, C = 1, W = 1000000, H = 1, G = 100000, V = 1000000$. Показаны «внутренние» ключевые слова.

Глава 7

Глобус

В этой главе описано, как для трехдольной полужесткой модели данных построить глобус.

7.1 Преимущества глобуса перед плоской картой

Обычная географическая карта существует в двух основных вариациях: в виде плоской карты и в виде глобуса. Каждая из них имеет свои преимущества и свои недостатки. Преимущества глобуса перед плоской картой связаны с тем, что он по форме ближе к планете Земля, чем плоская карта.

В географической карте площадные формы заданы извне. В задаче, поставленной в главе 5, формы нужно найти. Возникает вопрос: имеет ли глобус какие-либо преимущества перед плоской картой применительно к задаче из главы 5? Ответ: да, имеет. Дело в том, что у глобуса нет «краев», в отличие от прямоугольника, а значит нет и некоторых краевых эффектов. Если речь идет о глобусе, то не встает вопрос об интерпретации положения формы у края карты или в центре карты. К тому же, при построении диаграммы Вороного на плоскости некоторые ячейки получаются неограниченными. Так как вся карта должна помещаться в отведенный для нее прямоугольник, то с этим приходится бороться, «обрезая» ячейки так, чтобы они помещались на этом прямоугольнике полностью. То есть формы, оказавшиеся на краю карты, определяются в том числе и границами прямоугольника. Это вносит дополнительные искажения. Получается, что формы в центре плоской карты определяются более точно, чем те, что находятся на краю. Если бы областью построения карты была сфера, таких проблем бы не было.

Из всего вышеперечисленного следует вывод, что имеет смысл построить карту на сфере, то есть построить глобус.

7.2 Подход к построению глобуса

Набор требований для глобуса точно такой же, как набор требований к плоской карте из раздела 5.2, за исключением того, что областью построения карты стано-

вится сфера. Подход к построению карты тоже не меняется, а остается почти таким же, как в разделе 5.3. Карта также строится с помощью диаграммы Вороного, только теперь это диаграмма на сфере [13].

Определение 3. *Диаграмма Вороного конечного множества точек M на сфере представляет собой такое разбиение сферы, при котором каждая область этого разбиения (ячейка) образована множеством точек, более близких к одному из элементов множества M (центру ячейки, вершине), чем к любому другому элементу множества M.*

В качестве вершин ячеек для диаграммы Вороного используются вершины графа. В этом графе есть только рубрики и «пограничные» ключевые слова. Построение глобуса проходит в те же 3 этапа, описанные в разделе 5.3, но областью диаграммы для каждой компоненты связности является сфера, то есть каждая компонента связности является отдельной «планетой».

7.3 Формализация задачи для глобуса

Так как набор требований и подход к построению карты все те же, то можно сформулировать математическую постановку задачи для глобуса аналогично таковой для плоской карты. Для этого нужно оптимизируемый функционал для плоской карты адаптировать к сфере, то есть, используя расстояние на сфере, преопределить радиусы рубрик и ключевых слов.

7.3.1 Расстояние и площадь круга на сфере

Самое важное в данном случае отличие сферы от плоскости состоит в том, что расстояние между двумя точками на сфере определяется иначе, нежели на плоскости.

Определение 4. *Расстояние между двумя точками на сфере — это длина меньшей из двух дуг большой окружности, проходящей через эти точки.*

Определение 5. *Большая окружность сферы — это окружность, лежащая на сфере, центр которой совпадает с центром сферы.*

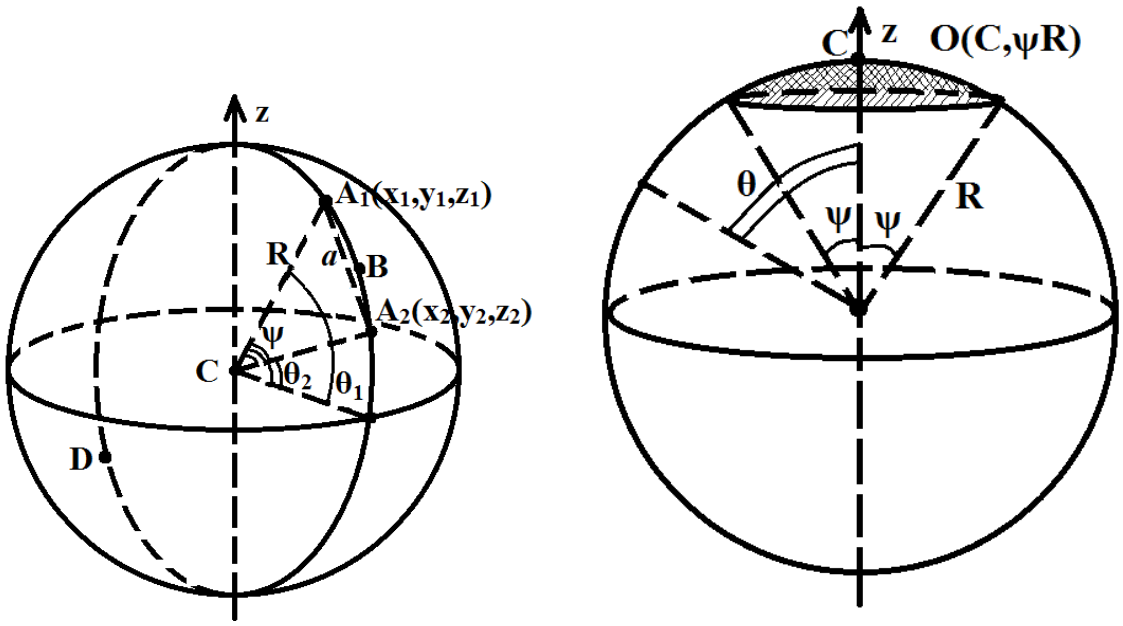
На рисунке (7.1(а)) расстояние между точками A_1 и A_2 равно

$$\rho(A_1, A_2) = \smile A_1BA_2 = \psi R,\ R - \text{радиус сферы},\ \psi = \angle A_1CA_2.$$

Так как $\rho(A_1, A_2) = \min(\smile A_1BA_2, \smile A_1DA_2)$ и $\smile A_1BA_2 + \smile A_1DA_2 = 2\pi R$, то

$$0 \leqslant \rho(A_1, A_2) \leqslant \pi R,\ 0 \leqslant \psi \leqslant \pi. \tag{7.1}$$

Тогда $\rho(A_1, A_2) = \arccos(\cos\psi)R$.



(a) (b)

Рис. 7.1: Расстояние и площадь на сфере

Из теоремы косинусов для треугольника $\triangle A_1CA_2$:

$$\cos\psi = \frac{A_1C^2 + CA_2{}^2 - A_1A_2{}^2}{2 \cdot A_1C \cdot CA_2} = \frac{2R^2 - a^2}{2R^2} = 1 - \frac{a^2}{2R^2},\ a = A_1A_2.$$

Переход к сферическим координатам:

$$\begin{cases} x = R\cos\theta\cos\varphi \\ y = R\cos\theta\sin\varphi \\ z = R\sin\theta \end{cases}, \text{ где } \theta \in [-\frac{\pi}{2}; \frac{\pi}{2}] \text{ и } \varphi \in [-\pi; \pi].$$

Тогда

$$a^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 = 2R^2(1 - \cos\theta_1\cos\theta_2\cos(\varphi_1 - \varphi_2) - \sin\theta_1\sin\theta_2),$$

и

$$\cos\psi = \cos\theta_1\cos\theta_2\cos(\varphi_1 - \varphi_2) + \sin\theta_1\sin\theta_2,$$

$$\psi = \arccos(\cos\theta_1\cos\theta_2\cos(\varphi_1 - \varphi_2) + \sin\theta_1\sin\theta_2), \tag{7.2}$$

$$\rho(A_1, A_2) = \arccos(\cos\theta_1\cos\theta_2\cos(\varphi_1 - \varphi_2) + \sin\theta_1\sin\theta_2)R. \tag{7.3}$$

Итак, получено выражение для расстояния между двумя точками на сфере. Теперь нужно найти площадь круга на сфере через его радиус.

Определение 6. *Круг на сфере $O(C,r)$ с центром в точке C и радиусом r — это множество точек A на сфере таких, что $\rho(C,A) \leqslant r$.*

Пусть есть круг $O(C,r)$. Пусть ось z проходит через центр круга — точку C, а угол θ отсчитывается от оси z, как показано на рисунке (7.1(b)), $\theta \in [0;\pi]$, $\varphi \in [0;2\pi]$. Тогда площадь круга $O(C,r)$ с радиусом $r = \psi R$, где R — радиус сферы, равна:

$$S = R^2 \int d\Omega = R^2 \int_0^{2\pi} d\varphi \int_0^{\psi} \sin\theta d\theta = 2\pi R^2 \int_0^{\psi} \sin\theta d\theta = -2\pi R^2 \cos\theta \,|_0^{\psi} = 2\pi R^2(1-\cos\psi).$$

Итого:

$$S(\psi) = 2\pi R^2(1-\cos\psi). \tag{7.4}$$

7.3.2 Постановка задачи для глобуса

Так как при фиксированном радиусе сферы R расстояние между двумя точками однозначно определяется углом ψ, то далее под расстоянием на сфере будет подразумеваться угловое расстояние:

$$\rho(A_1,A_2) = \psi = \arccos(\cos\theta_1\cos\theta_2\cos(\varphi_1-\varphi_2) + \sin\theta_1\sin\theta_2).$$

Тогда радиусы рубрик и ключевых слов тоже будут углами.

Угловое расстояние зависит только от двух координатных углов, поэтому далее под координатами рубрик и ключевых слов понимаются их координатные углы: $\mathbf{x}_i = (\theta_i, \varphi_i)$, и $\mathbf{y}_k = (\theta_k, \varphi_k)$ — координаты i-ого ключевого слова и k-ой рубрики соответственно.

Из условия пропорциональности площади формы количеству относящихся к соответствующей рубрике публикаций получается:

$$s_k = 2\pi R^2(1-\cos R_k) = \alpha d_k \Rightarrow \cos R_k = 1 - \frac{\alpha d_k}{2\pi R^2},$$

$$R_k = \arccos\left(1 - \frac{\alpha d_k}{2\pi R^2}\right), \tag{7.5}$$

где R — радиус сферы (области диаграммы), R_k — угловой радиус k-ой рубрики, d_k — количество публикаций, относящихся к k-ой рубрике, α — коэффициент пропорциональности. Из условия (7.1) следует, что

$$0 \leqslant \cos R_k \leqslant 1 \Rightarrow 0 \leqslant 1 - \frac{\alpha d_k}{2\pi R^2} \leqslant 1 \Rightarrow 0 \leqslant \alpha \leqslant \frac{2\pi R^2}{d_k}. \tag{7.6}$$

Формы должны покрывать всю сферу, следовательно

$$\sum_k s_k = \sum_k \alpha d_k = 4\pi R^2 \Rightarrow \alpha = \frac{4\pi R^2}{\sum_k d_k}. \tag{7.7}$$

Из (7.6) и (7.7) следует, что должно выполняться неравенство:

$$\alpha = \frac{4\pi R^2}{\sum_k d_k} \leqslant \frac{2\pi R^2}{\max_k d_k} \Rightarrow \max_k d_k \leqslant \frac{1}{2}\sum_k d_k.$$

С учетом (7.5) и (7.7) получается $R_k = \arccos\left(1 - \frac{2d_k}{\sum_i d_i}\right)$. Аналогично угловой радиус i-ого ключевого слова равен $r_i = \arccos\left(1 - \frac{2q_i}{\sum_i d_i}\right)$, где q_i — количество публикаций у i-ого ключевого слова, относящихся не только к той рубрике, к которой оно относится. Тогда оптимизационные задачи (5.1) и (5.5) для плоской карты и аналогичные им задачи (7.8) и (7.9) для глобуса записываются почти одинаково:

$$\sum_{i,j} \chi(\mathbf{y}_i, \mathbf{y}_j) \to \min, \tag{7.8}$$

$$W\sum_{k,i} f(\mathbf{y}_k, \mathbf{x}_i) + \sum_{i,j}\left(Hg_{ij}\rho(\mathbf{x}_i, \mathbf{x}_j)^2 + G\frac{r_i + r_j}{\rho(\mathbf{x}_i, \mathbf{x}_j)}\right) + V\sum_{i,j}\frac{R_i + R_j}{\rho(\mathbf{y}_i, \mathbf{y}_j)} \to \min. \tag{7.9}$$

Здесь функции $\chi(\mathbf{y}_i, \mathbf{y}_j)$ и $f(\mathbf{y}_k, \mathbf{x}_i)$ определяются соотношениями (5.2), (5.3) и (5.4).

7.4 Вычислительный эксперимент

Вычислительный эксперимент проводится на данных, описанных в главе 1.3. При этом использовался 3-ий подход к решению задачи оптимизации, описанный в главе 6, с небольшими дополнениями:

- В качестве начального приближения для координат рубрик берутся координаты центров соответствующих прямоугольников, построенных с помощью treemapping. Эти координаты преобразуются к координатным углам следующим способом:
$$\begin{cases} \varphi = \dfrac{2\pi(x - x_0 - \frac{w}{2})}{w} \\ \theta = \dfrac{2\pi(y - y_0 - \frac{h}{2})}{h} \end{cases},$$
где $\theta \in [-\frac{\pi}{2}; \frac{\pi}{2}]$ и $\varphi \in [-\pi; \pi]$, $x_0,\ y_0,\ w,\ h$ — координаты нижнего левого угла прямоугольника, построенного с помощью treemapping и соответствующего обрабатываемой компоненте связности, его ширина и высота соответственно.

- Решается оптимизационная задача (7.8), в процессе решения которой изменяются только координаты рубрик.

- Для каждой рубрики определяется наименьшее угловое расстояние до другой рубрики, оно делится пополам и внутри круга с таким радиусом располагаются пограничные ключевые слова этой рубрики.

- Все полученные координаты используются как начальное приближение для решения задачи (7.9), где меняются координаты и рубрик, и ключевых слов.
- Полученные угловые координаты преобразуются в декартовые:

$$\begin{cases} x = R\cos\theta\cos\varphi \\ y = R\cos\theta\sin\varphi \\ z = R\sin\theta \end{cases}, \text{ где } \theta \in [-\frac{\pi}{2};\frac{\pi}{2}] \text{ и } \varphi \in [-\pi;\pi].$$

На рисунках (7.2) и (7.3) приведен один и тот же глобус для самой большой компоненты связности с разных сторон. Этот результат получен при тех же значениях числовых параметров, что и лучший результат для плоской карты: $M = 3, L = 7.5, C = 1, W = 1000000, H = 1, G = 100000, V = 1000000$.

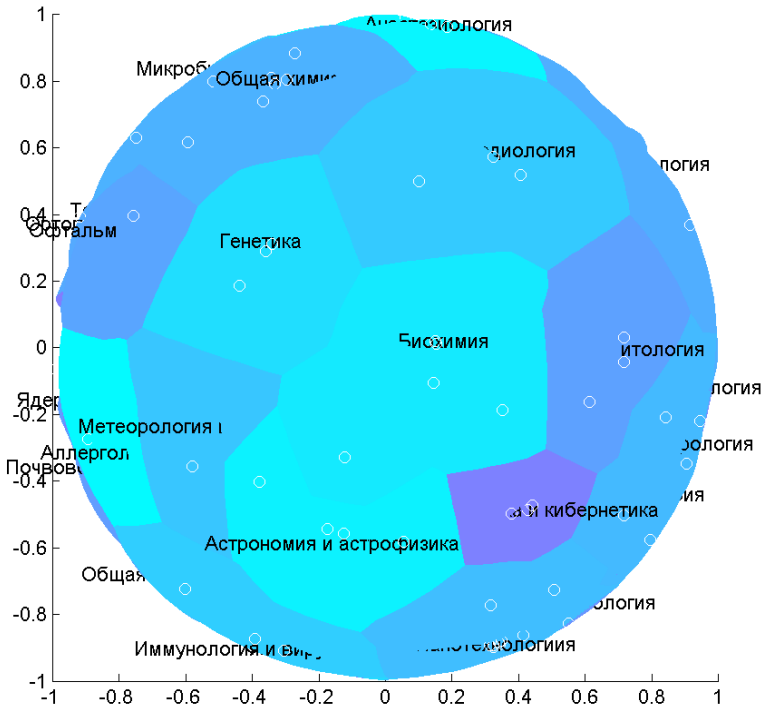


Рис. 7.2: Глобус для самой большой компоненты связности. Показаны все ключевые слова.

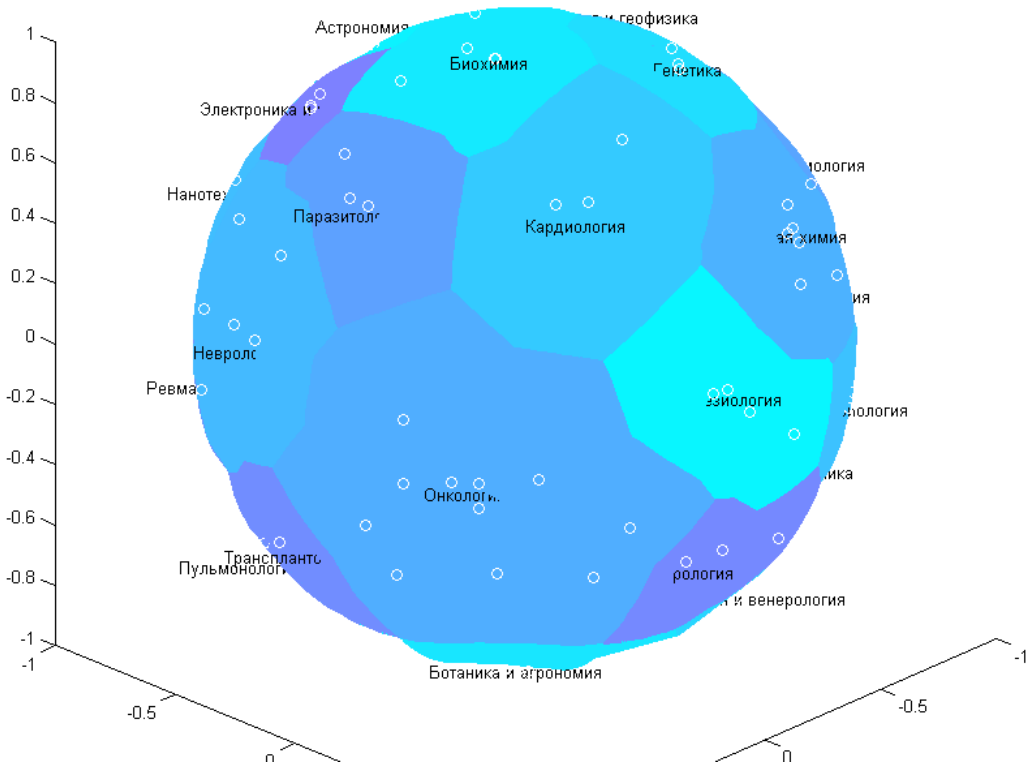


Рис. 7.3: Глобус для самой большой компоненты связности. Показаны все ключевые слова.

Заключение

Проведено исследование в соответствии с методикой разработки метода визуализации данных. Результатом работы является метод визуализации данных, описываемых трехдольной полужесткой моделью. С помощью этого метода данные могут быть представлены в виде плоской карты или в виде глобуса. Проведен вычислительный эксперимент на реальных данных.

Литература

[1] Препарата, Ф. Вычислительная геометрия: Введение / Ф. Препарата, М. Шеймос. — Мир, 1989.

[2] Andrienko, G. Choropleth maps: Classification revisited / Gennady Andrienko, Natalia Andrienko, Alexandr Savinov // ICC 2001: 6-10 August 2001. — 2001. — http://geoanalytics.net/and/papers/ica01.pdf.

[3] Automatically drawing euler diagrams with circles / Gem Stapleton, Jean Flower, Peter Rodgers, John Howse // Journal of Visual Languages and Computing. — 2012. — Vol. 23.

[4] Balzer, M. Voronoi treemaps / Michael Balzer, Oliver Deussen // IEEE Symposium on Information Visualization. — 2005.

[5] Chow, S. Drawing area-proportional venn and euler diagrams / Stirling Chow, Frank Ruskey // Graph Drawing, 11th International Symposium, GD 2003 Perugia, Italy, September 21-24, 2003 Revised Papers. — Springer Berlin Heidelberg, 2003. — P. 466–477.

[6] Engdahl, B. Ordered and unordered treemap algorithms and their applications on handheld devices. — 2005.

[7] Force-directed graph drawing using social gravity and scaling / Michael J. Bannister, David Eppstein, Michael T. Goodrich, Lowell Trott // Graph Drawing, 20th International Symposium, GD 2012, Redmond, WA, USA, September 19-21, 2012, Revised Selected Papers. — Springer Berlin Heidelberg, 2012. — P. 414–425.

[8] Gastner, M. T. Diffusion-based method for producing density-equalizing maps / M. T. Gastner, M. E. J. Newman // Proceedings of the National Academy of Sciences of the United States of America. — 2004. — Vol. 101. — P. 7499—-7504.

[9] Holten, D. Hierarchical edge bundles: visualization of adjacency relations in hierarchical data / Danny Holten // Visualization and Computer Graphics, IEEE Transactions on. — 2006. — Vol. 12. — P. 741–748.

[10] House, D. Continuous cartogram construction / D.H. House, C.J. Kocmoud // Visualization '98. Proceedings. — 1998. — P. 197–204.

[11] Johnson, B. S. Treemaps: Visualizing Hierarchical and Categorical Data: Ph.D. thesis / University of Maryland at College Park. — 1993.

[12] Kobourov, S. G. Spring embedders and force directed graph drawing algorithms. — 2012. — http://arxiv.org/pdf/1201.3011v1.pdf.

[13] Na, H.-S. Voronoi diagrams on the sphere / Hyeon-Suk Na, Chung-Nim Lee, Otfried Cheong // Computational Geometry. — 2002. — Vol. 23. — P. 183–194.

[14] Ruskey, F. A survey of venn diagrams / Frank Ruskey, Mark Weston // THE ELECTRONIC JOURNAL OF COMBINATORICS. — 2005. — http://www.combinatorics.org/files/Surveys/ds5/VennEJC.html.

[15] Schulz, H. Treevis.net: A tree visualization reference. — 2011. — http://vcg.informatik.uni-rostock.de/~hs162/treeposter/poster.html.

[16] Schulz, H.-J. The design space of implicit hierarchy visualization: A survey / Hans-Jorg Schulz, Steffen Hadlak, Heidrun Schumann // Visualization and Computer Graphics, IEEE Transactions on. — 2011. — Vol. 17. — P. 393–411.

[17] Schulz, H.-J. Force-directed lombardi-style graph drawing / Hans-Jorg Schulz, Steffen Hadlak, Heidrun Schumann // Graph Drawing, 19th International Symposium, GD 2011, Eindhoven, The Netherlands, September 21-23, 2011, Revised Selected Papers. — Springer Berlin Heidelberg, 2011. — P. 320–331.

[18] Vliegen, R. Visualizing business data with generalized treemaps / R. Vliegen, J.J. van Wijk, E.-J. van der Linden // Visualization and Computer Graphics, IEEE Transactions on. — 2006. — Vol. 12. — P. 789–796.

[19] Wilkinson, L. Exact and approximate area-proportional circular venn and euler diagrams / Leland Wilkinson // Visualization and Computer Graphics, IEEE Transactions on. — 2012. — Vol. 18. — P. 321–331.