Обзор средств визуализации тематических моделей коллекций текстовых документов*

Айсина Р. М.
rose.aysina@gmail.com
Московский государственный университет им. М. В. Ломоносова,
119991, Российская Федерация, Москва, Ленинские горы, д. 1

Тематическое моделирование является важным инструментом статистического анализа текстовых коллекций. Наглядное представление тематической модели позволяет лучше изучить кластерную структуру коллекции и оценить качество тематической модели. Средства визуализации являются неотъемлемой частью графических пользовательских интерфейсов, облегчающих тематический поиск и навигацию по коллекции. В обзоре описываются средства визуализации на основе веб-интерфейсов для иерархических, динамических и мультимодальных тематических моделей. Приводятся примеры визуализации графов и сетей, предлагается систематизация средств визуализации тематических моделей по их функциональным возможностям.

Ключевые слова: *анализ текстов, тематическое моделирование, кластеризация, научная визуализация, иерархия, граф, сеть.*

Survey of visualization tools for topic models of text corpora*

Aysina R. M.
Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation

Topic modeling is an important tool for statistical analysis of text collections. A visual representation of a topic model enables researchers to study cluster structure of the collection and estimate quality of the topic model. Visualization tools are especially important for graphical user interfaces as they facilitate search and navigation across documents of the collection. In this survey we describe web-based visualization tools for topic models, including hierarchical, temporal and multimodal models. We give examples of graph and network visualization, and categorize visualization tools according to their functionality.

Keywords: *text mining, topic modeling, clustering, scientific visualization, hierarchy, graph, network.*

Введение

Тематическое моделирование (*topic modeling*) — одно из направлений статистического анализа текстов, активно развивающееся с конца 90-х годов [3]. Тематическая модель коллекции текстовых документов определяет, какие термины (ключевые слова или словосочетания) образуют каждую тему, и какие темы образуют тематику каждого документа. Тематические модели применяются для выявления трендов в научных публикациях и новостных потоках, для классификации и категоризации документов, изображений и видео, в информационном поиске, в рекомендательных системах и в других приложениях.

Работа выполнена при поддержке Российского научного фонда (проект №15-18-00091).

Разработаны сотни специализированных моделей, учитывающих различные особенности текстов естественного языка и различные виды дополнительной информации [15]. Многомодальные модели учитывают метаданные документов и позволяют определять тематику не только самих документов, но и связанных с ними *объектов* различных *модальностей* — авторов, пользователей, тэгов, источников, категорий, классов, именованных сущностей, изображений, и т. д. Динамические модели учитывают время публикации документов и позволяют отслеживать изменения тематики документов и других объектов во времени. Иерархические модели строят иерархическую тематическую структуру, рекурсивно разделяя темы на подтемы. Сетевые модели учитывают взаимосвязи между документами посредством гиперссылок, цитирования, авторства или комментирования. Тематическое моделирование всё чаще применяется для выявления тематических сообществ в социальных сетях.

При таком богатстве моделей и приложений возникает потребность в средствах визуализации. Чем больше объём коллекции, тем острее стоит проблема наглядного представления как исходных данных, так и результатов тематического моделирования. Визуализация обычно преследует несколько целей одновременно. Как минимум, пользователям предоставляется возможность тематического поиска и тематической навигации по коллекции.

Тематический поиск — это возможность по документу, слову, объекту или теме найти документы, слова, объекты той же или схожей тематики. В тематическом поиске, в отличие от более привычного полнотекстового поиска, запросом может быть не только короткая текстовая строка, но и документ произвольной длины. Система тематического поиска определяет тематику документа и формирует результаты поиска либо в виде ранжированного списка, либо в виде структурированного графического представления.

Тематическая навигация — это возможность лёгкого (по одному клику) перехода пользователя от любого визуального элемента, представляющего документ, тему, объект, термин, и т. д. к тематически связанным с ним элементам, в частности, переход от документа к списку (или иному визуальному представлению) его тем, от темы — к списку релевантных ей документов, объектов, терминов, и т. д.

Простейшие средства визуализации непосредственно отображают результаты тематического моделирования — распределения тем для каждого документа и распределения терминов для каждой темы. Они могут отображаться в виде ранжированных списков, либо с помощью графических средств.

Более функционально богатые средства визуализации предоставляют различные способы отображения кластерной тематической структуры коллекции. Для этого могут использоваться графы, диаграммы, матрицы, сети. Цели визуализации понимаются разными исследовательскими группами и разработчиками по-разному. Это приводит к большому разнообразию идей визуализации, от выбора отображаемых структурных особенностей тематической модели до выбора элементов графического дизайна.

Некоторые средства визуализации нацелены на упрощение экспертной оценки качества (интерпретируемости) тематической модели, и даже позволяют вмешиваться в процесс построения модели, изменяя, удаляя или добавляя темы в документах или термины в темах.

В данном обзоре описаны основные идеи визуализации тематических моделей на основе веб-интерфейсов. Рассмотрены визуализаторы для плоских моделей (*Topic Model Visualization Engine, Termite System и TopicNets*), для иерархических моделей (*Hiérarchie, iVisClustering*), для динамических моделей (*TextFlow*), для иерархических динамических моделей (*HierarchicalTopics* и *RoseRiver*). В заключении приводится систематизация средств визуализации по их функциональным возможностям.

Вероятностное тематическое моделирование

Пусть D — множество (*коллекция*) текстовых документов, W — множество (*словарь*) всех употребляемых в них терминов (слов или словосочетаний). Каждый документ $d \in D$ представляет собой последовательность n_d терминов $w_1, \ldots, w_{n_d}$ из словаря W.

Вероятностная тематическая модель (ВТМ) описывает каждый документ d дискретным распределением $p(t \mid d)$ на множестве тем T, каждую тему $t \in T$ — дискретным распределением $p(w \mid t)$ на множестве терминов W. Можно также говорить о совместной *«мягкой» кластеризации* множества документов и множества слов по множеству кластеров-тем. «Мягкая» кластеризация означает, что каждый документ или термин не жёстко приписывается какой-то одной теме, а распределяется по нескольким темам.

Множество тем T чаще всего задаётся как конечное множество заданной мощности. Модель сама определяет, какие слова с какими вероятностями войдут в каждую тему. Это создаёт *проблему интерпретируемости тем*. Тематическая модель не гарантирует, что каждая тема будет иметь содержательную интерпретацию с точки зрения людей — экспертов, понимающих тематику данной коллекции. Тема считается интерпретируемой, если по ранжированному списку наиболее релевантных слов данной темы эксперт в состоянии понять, о чём эта тема, и дать ей название [7]. При визуализации тематических моделей, построенных автоматически без участия экспертов, возникает проблема автоматического именования тем. В простейшем случае она решается путём конкатенации нескольких наиболее репрезентативных слов темы [8]. Другие подходы к автоматическому именованию тем можно найти в [27, 23, 5].

Для построения распределений тем в документах и слов в темах используются различные модели и методы. Самыми простыми являются модели *вероятностного латентного семантического анализа* PLSA (Probabilistic Latent Semantic Analysis) [21] и *латентного размещения Дирихле* LDA (Latent Dirichlet Allocation) [2]. Подавляющее большинство тематических моделей, разработанных за последние 15 лет, являются их модификациями [15]. Несмотря на различные усложнения, в большинстве моделей на выходе формируются те же структуры данных — распределения терминов в темах и тем в документах. Поэтому базовые средства визуализации этих распределений могут быть применены к большинству моделей.

На выходе более сложных моделей могут формироваться *тематические профили* $p(t \mid x)$ объектов x различных модальностей — авторов, моментов времени, категорий, и т. д. Иногда их пересчитывают по формуле Байеса в распределения $p(x \mid t) = p(t \mid x)\frac{p(x)}{p(t)}$, чтобы показывать наиболее релевантные объекты в темах t. Распределения $p(x)$ и $p(t)$ легко оцениваются в процессе построения модели. Для отображения таких данных и их взаимосвязей разрабатываются специализированные средства визуализации.

Система TMVE

Система *Topic Model Visualization Engine* (TMVE) [8] — это навигатор по коллекции, имеющий два основных типа страниц: страница темы и страница документа. Есть также обзорные страницы, отображающие общую структуру коллекции. Они являются стартовыми, с них начинается работа с навигатором, рис. 1.

Страница темы разделена на три колонки. Слева находится список слов w, упорядоченных по убыванию вероятности $p(w \mid t)$ в данной теме t. По этой последовательности пользователь обычно может быстро понять, о чём тема. Названия тем формируются автоматически как тройки наиболее представительных слов. В центре находится список документов, упорядоченных по убыванию вероятности $p(d \mid t)$. Справа располагается список

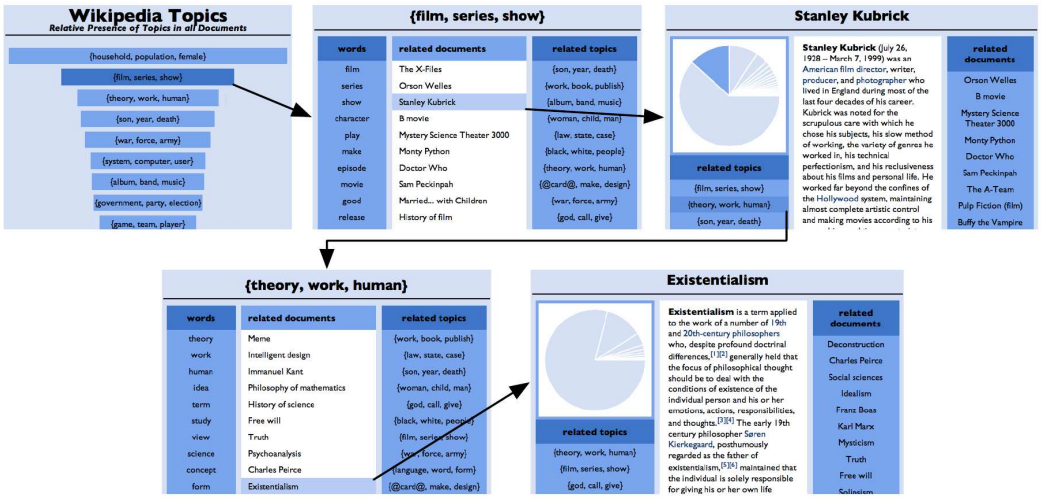

Рис. 1. Система *TMVE*. Навигация по Википедии. С разрешения авторов [8].

Рис. 2. Система *TMVE*. Страница документа (слева) и страница темы (справа). С разрешения авторов [8].

схожих тем, имеющих наиболее близкие распределения слов. Для оценивания сходства тем s и t используется функция расстояния

$$R(s,t) = \sum_{w \in W} [p(w \mid s) > 0]\,[p(w \mid t) > 0]\,\bigl|\log p(w \mid t) - \log p(w \mid s)\bigr|.$$

Эта функция подходит для модели LDA [2], в которой вероятности $p(w \mid t)$, как правило, отличны от нуля. Однако для сравнения тем в разреженных моделях лучше использовать расстояние Хеллингера или косинусное расстояние, не содержащие логарифмов.

Страница документа. В центре отображается текст документа, слева от него — темы, к которым относится документ, и секторная диаграмма, показывающая вероятности $p(t \mid d)$ тем в данном документе. Слева находится список документов, имеющих ту же тематику, что и данный документ (рис. 2).

Обзорные страницы являются «точками входа» для навигации по коллекции. На них представлены темы, упорядоченные согласно вероятностям $p(t)$, причём размер поля пропорционален вероятности (рис. 3).

Рассмотрим взаимодействие пользователя с системой на примере визуализации статей Википедии. Вначале пользователь видит набор тем, составляющих коллекцию. Выбрав тему {*film, series, show*}, пользователь попадает на страницу документов этой темы. Затем, выбрав документ «*Stanley Kubrick*», можно просмотреть саму статью и темы, к кото-

Рис. 3. Система *TMVE*. Обзорная страница. С разрешения авторов [8].

рым она относится. При выборе схожей темы {*theory, work, human*} пользователь переходит на страницы документов уже этой темы, где он, например, может прочитать статью «*Existentialism*».

Достоинства: удобный интерфейс для навигации по коллекции; имеются списки схожих документов и схожих тем; возможность просмотра любого документа; автоматическое именование тем; открытый исходный код; возможность адаптации кода для конкретной задачи (возможно создание пользовательских режимов, изменение вида входных данных, изменение алгоритма построения тематической модели). Разработчики прилагают три демонстрационные версии на различных коллекциях (включая Википедию) с исходным кодом, на основе которого можно создавать свои браузеры.

Недостатки: названия тем из трех слов не всегда адекватны; нет визуализации других модальностей, кроме слов; нет возможности изменить модель.

Ссылки:
`https://code.google.com/p/tmve` — сайт проекта;
`http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html` — пример визуализации.

Система Termite

Система *Termite* [10] позволяет визуализировать матрицу терминов тем $p(w \mid t)$ и сравнивать темы друг с другом. Значения в матрице отображаются в виде кругов, радиусы которых пропорциональны вероятностям терминов в темах $p(w \mid t)$. Круги могут накладываться друг на друга. Пользователь может перейти к теме, нажав на круг или на название темы в матрице. При этом раскрываются два дополнительных представления темы. Первое — вероятности слов темы относительно всей коллекции, второе — документы, принадлежащие этой теме (рис. 4).

Termite может фильтровать термины, чтобы показать наиболее вероятные или значимые (*salient*) термины (рис. 5). Пользователь задает число отображаемых терминов от 10 до 250. Список самых вероятных слов содержит общие слова (*based, paper, approach*), в то время как список значимых слов, получаемый с помощью т. н. меры значимости (*saliency measure*), содержит слова, которые характерны для данной темы (*tree, context, task*).

Для определения меры значимости термина w вычисляется условная вероятность $p(t \mid w)$ и вероятность $p(t)$. Отличительность (*distinctiveness*) термина w определяется дивергенцией Кульбака-Лейблера между $p(t \mid w)$ и $p(t)$:

$$distinctiveness(w) = \sum_{t \in T} p(t \mid w) \log \frac{p(t \mid w)}{p(t)}.$$

Значимость (*saliency*) термина w определяется следующей формулой:

$$saliency(w) = p(w) \times distinctiveness(w).$$

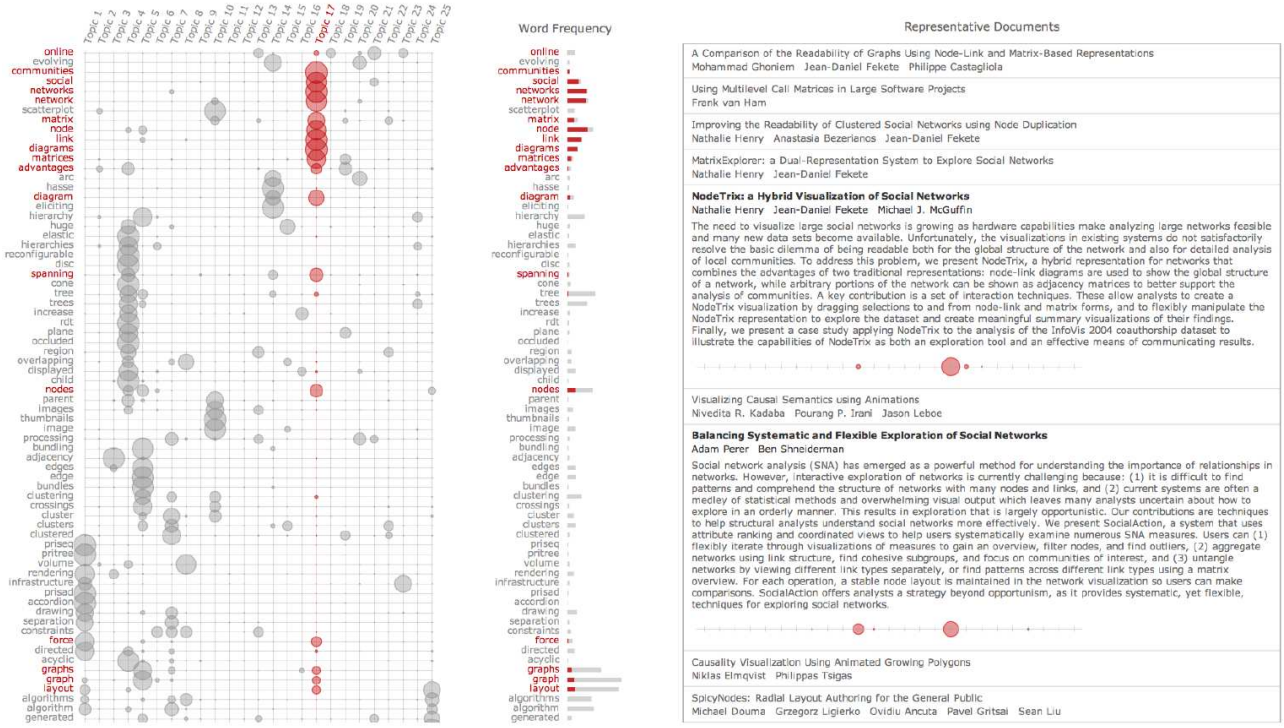

Рис. 4. Система *Termite*. Когда тема выбрана в матрице (слева), система отображает распределение терминов в теме и во всей коллекции (в середине) и показывает документы, наиболее соответствующие выбранной теме (справа). С разрешения авторов [10].

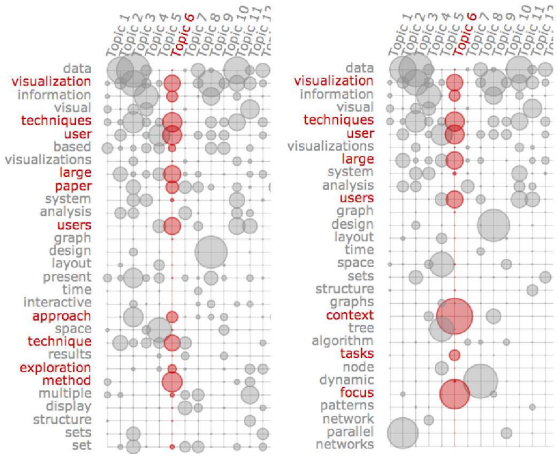

Рис. 5. Система *Termite*. Первые 30 частых (слева) и существенных (справа) терминов. С разрешения авторов [10].

Если генерировать более разреженную матрицу, то оценка значимости терминов обеспечивает более быструю дифференциацию тем и выявление потенциальных «ненужных» тем, в которых мало значимых терминов.

Termite также предлагает три опции для ранжирования терминов: по алфавиту, по частоте и по совместной встречаемости между парами соседних слов. Для ранжирования по совместной встречаемости оценивается вероятность того, что два слова неслучайно часто появляются вместе. Например, «*social network*» — более вероятная фраза, чем «*network social*» (рис. 6). Такое ранжирование улучшает интерпретируемость тем. Например, в те-

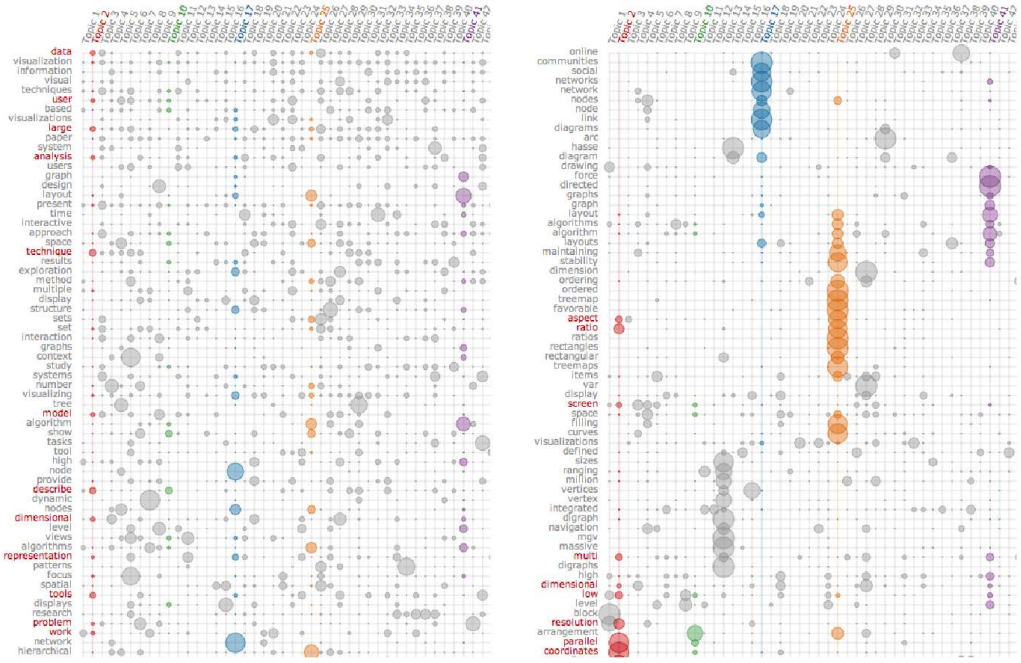

Рис. 6. Система *Termite*. Термины могут ранжироваться по частоте (слева), либо по совместной встречаемости (справа). С разрешения авторов [10].

ме 6 существенными словами являются *focus* и *context* (рис. 6, слева), в то время как самыми частыми словами являются общие слова *technique* и *method*. В теме 2 характерными являются словосочетания *aspect ratio* и *parallel coordinates* (рис. 6, справа).

Достоинства: несколько способов оценить интерпретируемость тем; различные виды представлений для тем; удобный интерфейс; открытый исходный код, адаптируемый под конкретную задачу.

Недостатки: нет визуализации распределения тем в документах; нет возможности внесения исправлений в модель при обнаружении плохой интерпретируемости; нет автоматического именования тем; нет возможности использовать термины-словосочетания.

Ссылки:
`https://github.com/uwdata/termite-visualizations.git` — сайт проекта;
`http://vis.stanford.edu/topic-diagnostics` — пример визуализации.

Система TopicNets

Система *TopicNets* [18] предназначена для визуализации и интерактивного анализа больших коллекций документов через веб-интерфейс. *TopicNets* представляет документы и темы вершинами графа. Главным достоинством *TopicNets* является поддержка интерактивного тематического моделирования. Модель перестраивается в режиме реального времени, прямо во время визуализации, для лучшего представления подмножеств тем и документов. Для этого используется распределенная реализация алгоритма свёрнутой вариационной байесовской аппроксимации *CVB0* (*Collapsed Variational Bayes*) [1]. Документы распределяются на несколько процессоров, и на каждом из них выполняются шаги *CVB0*, которые потом синхронизируются между собой.

Создание графа «документы-темы». Первым шагом создания графа является тематическое моделирование документов d, в результате которого вычисляются распределения тем в документах $p(t \mid d)$. Если $p(t \mid d)$ превышает установленный пользователем порог, то

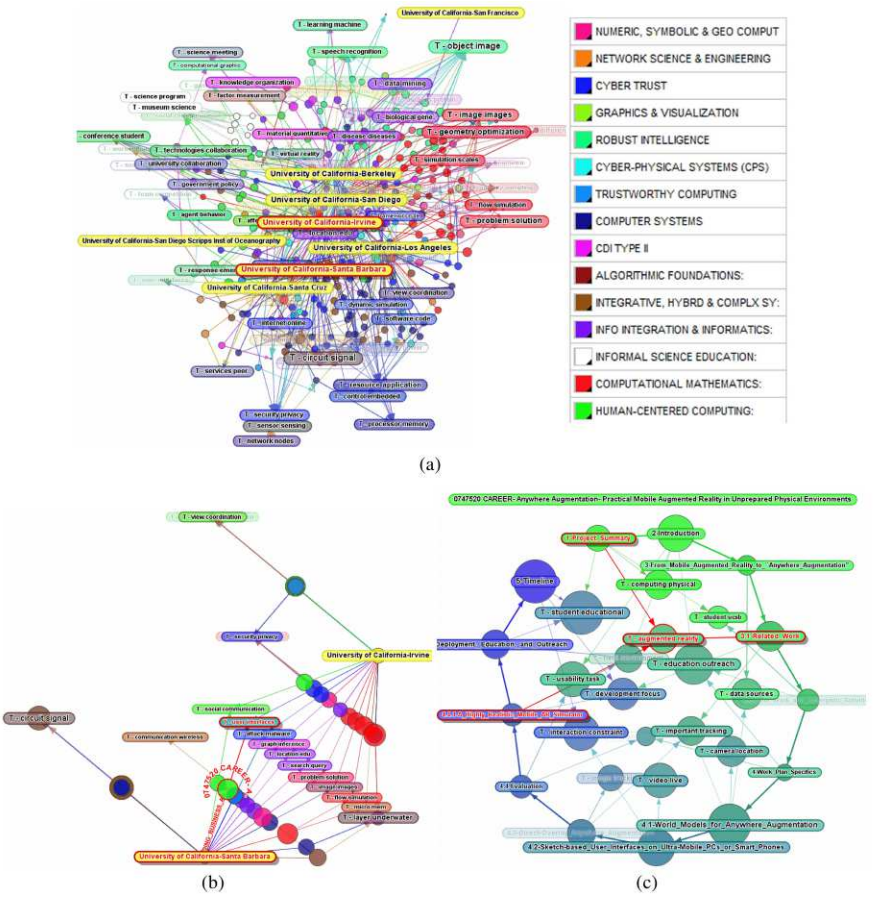

Рис. 7. Система *TopicNets*. Коллекция грантов *NSF*. Выбраны вершины, связанные с университетами. Вершины-темы обозначены буквой «T» и раскрашены согласно связанным с ними документам. (a) Гранты по тематике *Computer Science*, полученные в университетах Калифорнии. (b) То же самое после того, как пользователь выбрал фрагмент графа и один отдельный грант. (c) Визуализация одного документа. Вершины, обозначающие секции документа, находятся по периметру, вершины тем — внутри фигуры. На рисунке выделена одна тема и все связанные с ней вершины. С разрешения авторов [18].

в графе документ d и тема t соединяются ребром, толщина которого пропорциональна $p(t \mid d)$. Далее вершины-темы именуются первыми n наиболее вероятными словами темы (число n также устанавливается пользователем), размер вершины-темы пропорционален вероятности $p(t)$ появления темы в коллекции (рис. 7). Размер вершины-документа пропорционален длине документа.

Раскраска вершин и ребер графа. Цвет вершин-документов определяется набором цветов, который задаётся пользователем или формируется на основе метаданных коллекции (при их наличии). Цвета можно определить для каждого автора (автоматически или вручную), тогда вершины документов будут наследовать цвета своих авторов. При наличии нескольких модальностей, например, авторов, времени создания, организаций и т. д., цвета смешиваются. Цвета интерполируются между последовательными вершинами на временной шкале (рис. 9b). Интерполяция происходит в RGB пространстве, что не даёт идеального результата для каждой цветовой пары, однако пользователь может выбрать, какие два цвета смешивать, чтобы корректно отобразить временную шкалу.

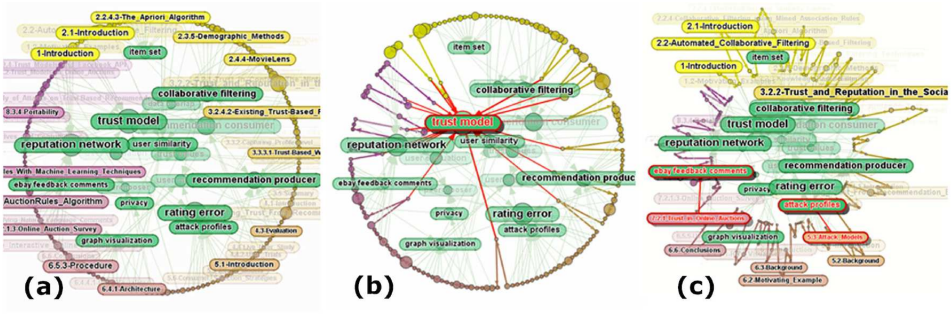

Рис. 8. Система *TopicNets*. (a) Линейная структура одного документа. (b) Распределение выделенной темы по документу. (c) Размер и расположение относительно центра окружности показывают близость каждой секции и темы к главной теме документа. С разрешения авторов [18].

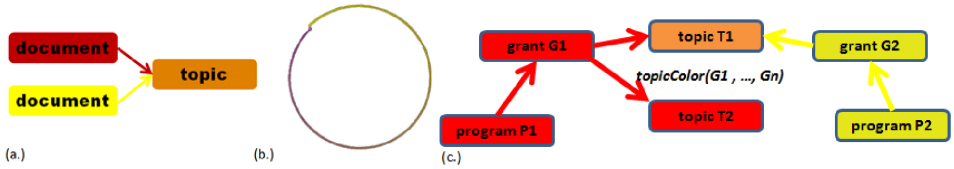

Рис. 9. Система *TopicNets*. (a) Пример смешения цветов для раскраски вершин-тем. (b) Пример интерполирования цвета для изображения временной шкалы. (c) Пользовательское задание цветов для раскраски вершин-тем. С разрешения авторов [18].

Ребра графа наследуют цвет вершины-документа, из которой они исходят. Цвет вершины-темы формируется путём смешения цветов ребер, входящих в эту вершину (рис. 9). Пользователь также имеет возможность задать свой набор цветов для тем (рис. 8).

Расположение схожих тем. Для определения сходства тем в *TopicNets* используется симметризованная дивергенция Кульбака–Лейблера между каждой парой распределений слов по темам $p(w \mid t)$. Результирующая матрица несходства тем является входными данными для алгоритма многомерного шкалирования (*multidimensional scaling*), который определяет позицию для каждой вершины-темы. Эти вершины затем фиксируются в соответствующей позиции, и применяется стандартный силовой алгоритм [16] для расстановки вершин-документов в пространстве соответствующей вершины-темы (при этом в качестве расстояния между темой и документом используется вероятность появления данной темы в данном документе).

Ранжирование документов. Если документы ранжированы по дате публикации, *TopicNets* расставляет их вершины по окружности (рис. 10). Это имеет определённые преимущества перед более привычным отображением временной шкалы в виде прямой линии. Вершина-тема может соединяться с большим числом документов, расположенных далеко друг от друга на прямой. Если тем много, изображение становится запутанным. Окружность сделана слегка винтообразной, чтобы вершины первого и последнего документа не встретились в одной точке.

Кроме сохранения хронологического порядка, алгоритм старается расположить схожие по тематике вершины близко к друг другу. В результате темы, которые появлялись в конкретный период времени, оказываются ближе к окружности (в секторе, соответствующем этому периоду), в то время, как темы, которые появлялись постоянно, располагаются ближе к центру. Если тема появляется в документах, расположенных диаметрально на окружности, то она также будет находиться в центре, что может ввести пользовате-

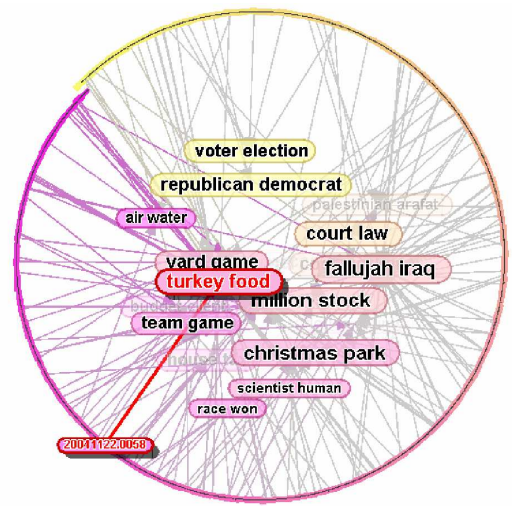

Рис. 10. Система *TopicNets*. Пример графа в *TopicNets*, показывающего темы новостей *NY Times* за ноябрь 2004. С разрешения авторов [18].

ля в заблуждение. Однако если эта тема появлялась редко, то размер её вершины будет маленьким. Таким образом, вершина-тема большого размера и близкая к центру может считаться наиболее релевантной для всей коллекции.

Фильтрация графа. В большинстве случаев пользователя интересует не весь граф тематической модели, а только его фрагмент. *TopicNets* позволяет задавать нужный фрагмент с помощью поискового запроса по названиям вершин или по наиболее вероятным словам тем. Затем можно выбрать нужные из найденных вершин и визуализировать только связанные с ними вершины, скрыв (по желанию) остальной граф. При этом система плавно трансформирует старый граф в новый. Щелчком мыши пользователь может перейти в другую часть графа или вернуться к предыдущей визуализации. Пользователь может выбрать нужные вершины в графе и скрыть другие вершины, не связанные с данными.

TopicNets позволяет добавлять различные типы метаданных и визуализировать их на исходном графе. Например, если в коллекции имеется информация об авторах документов, то возможно свёртывание вершин-документов в новые вершины авторов. При этом строится новый граф, соединяющий авторов и темы, и все принципы, описанные выше, сохраняются и для этого графа.

Визуализация одного документа. Все перечисленные способы визуализации могут быть применены не только к ранжированному множеству документов, но и их содержимому одного отдельного документа.

Веб-архитектура. Многие приложения для визуализации тематических моделей, способные сгенерировать интерактивный граф с высокой степенью масштабируемости, устанавливаются как отдельное приложение или плагин для веб-браузера, что может быть ресурсоёмким для клиентской машины. *TopicNets* построен на архитектуре *WiGis*, разработанной авторами *TopicNets* для визуализации графов на основе *AJAX*[1]. Это позволяет запускать *TopicNets* из веб-браузера и масштабировать граф, состоящий из сотен тысяч

[1] AJAX (*Asynchronous Javascript and XML*) — технология построения интерактивных пользовательских интерфейсов для веб-приложений, поддерживающая фоновый обмен данными между браузером и веб-сервером и постепенную загрузку веб-страниц.

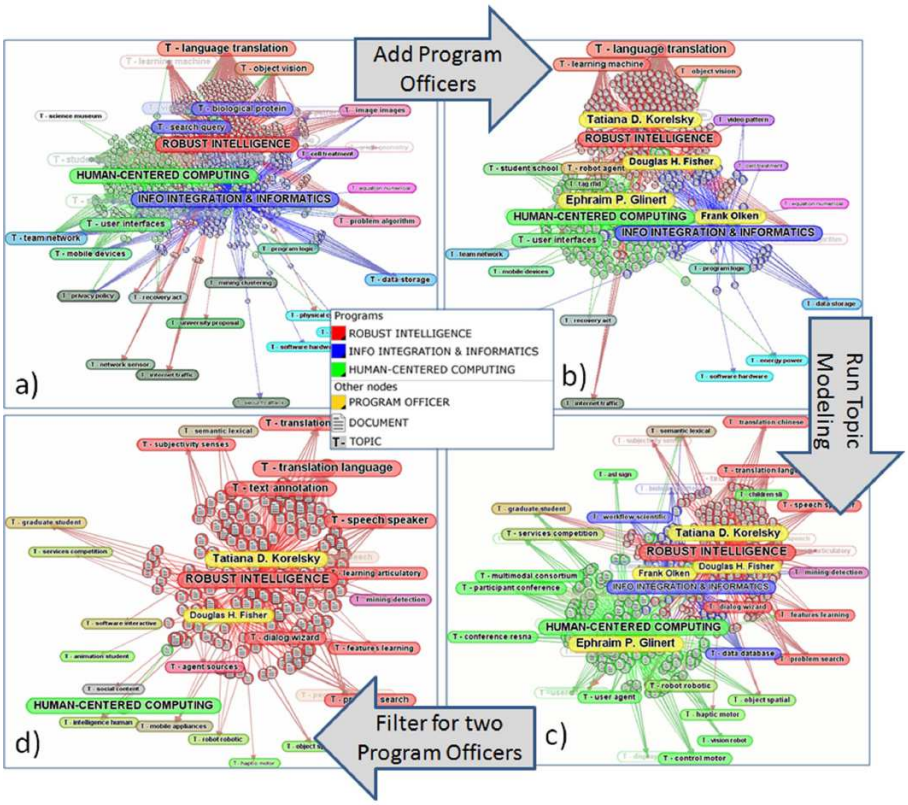

Рис. 11. Система *TopicNets*. Коллекция грантов *NSF*. С разрешения авторов [18].

документов. Все вычисления и формирование новых изображений для графа происходят на удаленном сервере.

Рассмотрим пример работы пользователя с коллекцией грантов *NSF*[2] (рис. 11). Сначала пользователь выбирает три темы, соответствующие программам фонда (рис. 11a), затем добавляет авторов — руководителей грантов (рис. 11b) и запускает повторное моделирование для подграфа, связанного только с авторами *Fisher, Korelsky, Glinert, Olken*. Затем он удаляет темы, не связанные с выбранным подграфом (*processor memory*), и добавляет новые (*language natural*) (рис. 11c). Далее он выбирает авторов *Fisher* и *Korelsky*, т. к. их вершины близко расположены друг к другу. На рис. 11d становится видно, что эти руководители занимаются программой *Robust Intelligence*, при этом тематика работ *Fisher* также пересекается с *Human-Centered Computing*, т. к. его вершина соединена с вершинами зеленого цвета.

Достоинства: богатый дружественный интерфейс; интеракивность; возможность учета модальностей; единые способы визуализации коллекции и отдельного документа; возможность масштабирования визуализации и уточнения тематической модели в режиме реального времени; запуск непосредственно из веб-браузера.

Недостатки: при большом числе тем и документов круговая визуализация временной шкалы становится неадекватной; сложность установки; необходимость устанавливать сторонние приложения; сложно адаптируемый код.

Ссылки:

`https://code.google.com/p/topicnets` — страница проекта *TopicNets*;

[2]Коллекция доступна на `http://www.nsf.gov/awardsearch/`

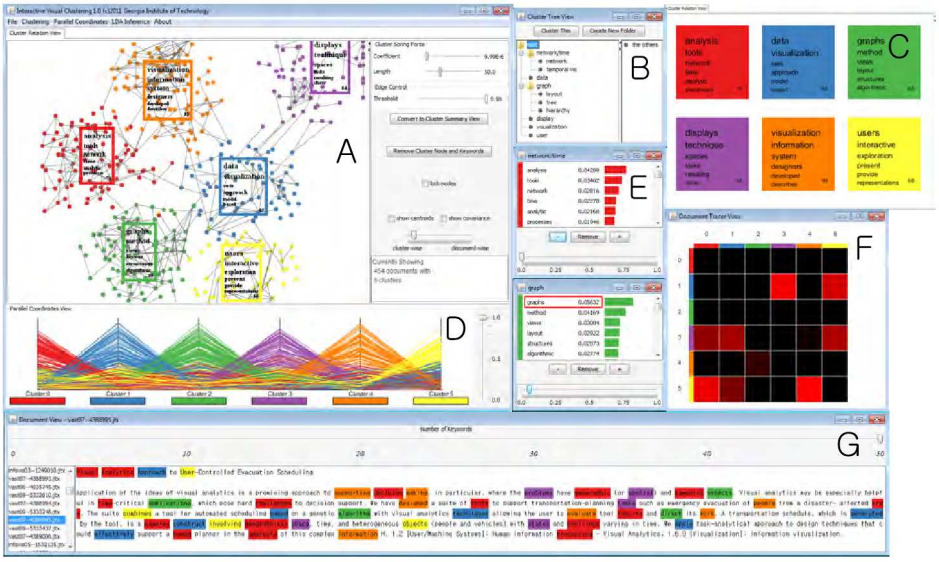

Рис. 12. Система *iVisClustering*. (A) *Cluster Relation View* — представление кластеров в виде графа. (B) *Cluster Tree View* — представление иерархии в виде дерева. (C) *Cluster Summary View* — представление кластеров без документов и связей между ними. (D) *Parallel Coordinates View* отображает темы, из которых состоят документы. (E) *Term-Weight View* — визуализация распределений $p(w \mid t)$ для каждой темы. (F) *Document Tracer View* — тепловая карта, показывающая, какие документы перешли из одного кластера в другой. (G) *Document View* позволяет просматривать отдельные документы. С разрешения авторов [24].

`https://code.google.com/p/wigis` — страница проекта *WiGis*;
`http://youtu.be/-Sgq-msjd-Y` — пример визуализации *TopicNets*.

Система iVisClustering

Система *iVisClustering* [24] позволяет полностью контролировать процесс кластеризации коллекции документов. Пользователь может создавать и удалять кластеры, разделять кластеры на более мелкие, производить повторную кластеризацию.

Система имеет несколько модулей для визуализации кластерной структуры коллекции.

Модуль *Cluster Relation View* представляет результаты кластеризации в виде графа, вершины которого соответствуют документам (рис. 12A). Для плоского размещения вершин графа используется силовой алгоритм многомерного шкалирования. Каждая вершина-документ изображается цветным кругом, документы одинакового цвета принадлежат одному кластеру. Длина ребра между двумя вершинами пропорциональна оценке сходства документов. Пользователь может задать пороговое значение сходства для отображения ребер: ребра, длина которых меньше этого значения, не будут отображаться.

Ключевые слова кластера изображаются в «общей вершине», которая представляет собой прямоугольник с цветной границей. Если при большом числе кластеров этот вид оказался перегруженным, то доступно еще одно представление — *Cluster Summary View*, где «общие вершины» отображаются в виде таблицы (рис. 12C).

Модуль *Parallel Coordinates View*. В этом представлении каждая вертикальная ось соответствует теме, а каждая линия обозначает документ (рис. 12D). Цвет линии соответствует цвету кластера, которому принадлежит документ. По вертикальной оси откладываются значения $p(t \mid d)$ для каждого документа d. Если линия документа имеет несколько

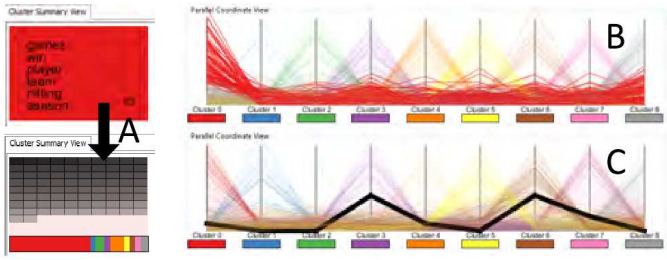

Рис. 13. *X-ray режим.* (A) Изображение таблицы документов, под таблицей находится цветовой спектр, показывающий, из каких тем состоит документ. (B) *Parallel Coordinates View* для кластера. (C) *Parallel Coordinates View* для одного документа, он изображен толстой черной линией. С разрешения авторов [24].

«пиков», то это значит, что он принадлежит нескольким темам. Пользователь может задать пороговое значение вероятности $p(t \mid d)$, ниже которого линии не отображаются.

Режим отображения *X-ray* активируется при наведении курсора на «общую вершину» кластера (рис. 13A), при этом документы кластера выделяются в *Parallel Coordinates View* (рис. 13B). Режим *X-ray* представляет документы кластера в виде таблицы, каждая ячейка которой соответствует документу. Чем выше вероятность $p(t \mid d)$, тем темнее ячейка. Если курсор наведен на ячейку, то соответствующий документ отображается в *Parallel Coordinates View* в виде толстой черной линии (рис. 13C). В режиме *X-ray* под таблицей документов находится цветной спектр: каждый цвет спектра соответствует кластеру, так что принадлежность выбранного документа тому или иному кластеру может быть определена визуально по доле цветной области в спектре.

Модуль *Term-Weight View* отображает распределение $p(w \mid t)$ всех терминов в теме (рис. 12E). Пользователь может изменять значения $p(w \mid t)$, после чего новая тематическая модель будет построена с учётом этих изменений. Например, если для термина w указанное значение увеличилось, то система начнёт чаще относить документы, содержащие w, к теме t. Документы, которые в результате переместились из одного кластера в другой, отображаются в *Document Tracer View*.

Модуль *Document Tracer View* отображает тепловую карту, показывающую перемещения документов между кластерами в процессе взаимодействия пользователя с системой (рис. 12F). Тепловая карта имеет размер $T \times T$, где T — количество кластеров. Каждый её элемент (i, j) отображает, как много документов перешли из кластера i в кластер j. При нажатии на элемент карты открываются соответствующие этому элементу документы в *Document View*.

Модуль *Document View* показывает отдельный документ (рис. 12G). При просмотре документа термины окрашиваются в цвета своих тем. Это представление доступно из любого модуля визуализации, кроме *Term-Weight View*.

Модуль *Cluster Tree View*. Одной из целей создания *iVisClustering* была возможность улучшения тематической модели. В этом представлении пользователь в процессе взаимодействия с системой может изменять иерархическую структуру модели. Для этого используются пять операций: разъединение и соединение кластеров, перемещение и удаление кластера, повторная кластеризация. Результаты этих действий немедленно отображаются в дереве иерархии (рис. 12B). Документы удаленного кластера сохраняются, так как они могут понадобиться при отыскании новых тем в коллекции, и их можно включить обратно в коллекцию в любой момент.

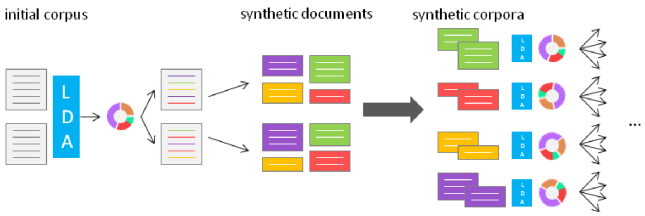

Рис. 14. Система *Hiérarchie*. Схема алгоритма *HLDA*. С разрешения авторов [29].

Повторная кластеризация используется при изменении числа кластеров. После неё с помощью *венгерского алгоритма* [22] находятся наилучшие парные соответствия между исходными и новыми кластерами. Цвета новых кластеров изменяются так, чтобы схожие кластеры до и после изменения имели одинаковый цвет.

Достоинства: интерактивное взаимодействие с пользователем с целью улучшения визуального представления и улучшения тематической модели; наличие нескольких различных представлений для отображения коллекции; возможность автоматического и ручного именования тем.

Недостатки: нет автоматического построения иерархической структуры (пользователь должен сам разбивать темы на подтемы); нет возможности тематического поиска.

Пример визуализации:
`ftp://temp:temp@mimi.cc.gt.atl.ga.us/resource/2012_eurovis_ivisclustering.mp4`

Система Hiérarchie

Система *Hiérarchie* [29] — это веб-приложение для построения иерархических тематических моделей и их визуализации в виде дерева.

Для построения тематической модели в *Hiérarchie* используется иерархический алгоритм *HLDA* (*Hierarchical latent Dirichlet allocation*), который рекурсивно разделяет темы на подтемы. Используя распределения $p(t \mid d)$, *HLDA* разделяет каждый документ d на *искусственные поддокументы* по каждой теме t. Поддокументы содержат только те слова, которые относятся к данной теме. Таким образом, для каждой темы t генерируется новая *искусственная подколлекция*, и для неё строится тематическая модель следующего уровня, разделяющая данную тему на подтемы (рис. 14). Процесс продолжается до тех пор, пока подколлекции не станут слишком маленькими для моделирования. Пользователь может задавать критерии остановки рекурсивного процесса.

В качестве основной визуализации в *Hiérarchie* используется круговая диаграмма с исходящими из нее лучами (*sunburst chart*). Она позволяет отображать как маленькие, так и большие иерархии без интерактивной прокрутки, приспосабливаясь к размеру экрана без искажения структуры иерархии.

Рис. 15 показывает верхний уровень для коллекции твитов и новостей, касающихся пропавшего самолета Малазийских авиалиний. Каждый уровень иерархии изображается в виде кольца, разбитого дуги, изображающие темы. Чтобы избежать загромождения диаграммы, темы не подписываются, вместо этого их названия изображаются сверху при наведении на них курсора. Таким образом, *Hiérarchie* придерживается принципа «сначала общий вид, затем масштабирование и фильтрация, детали по запросу». При наведении мыши на тему в середине схемы отображаются слова темы, что экономит пространство экрана и требует минимальных перемещений взгляда.

Когда пользователь выбирает тему, диаграмма перестраивается для отображения только выбранной темы и её подтем (рис. 15). Сверху строится путь от корня всей иерархии

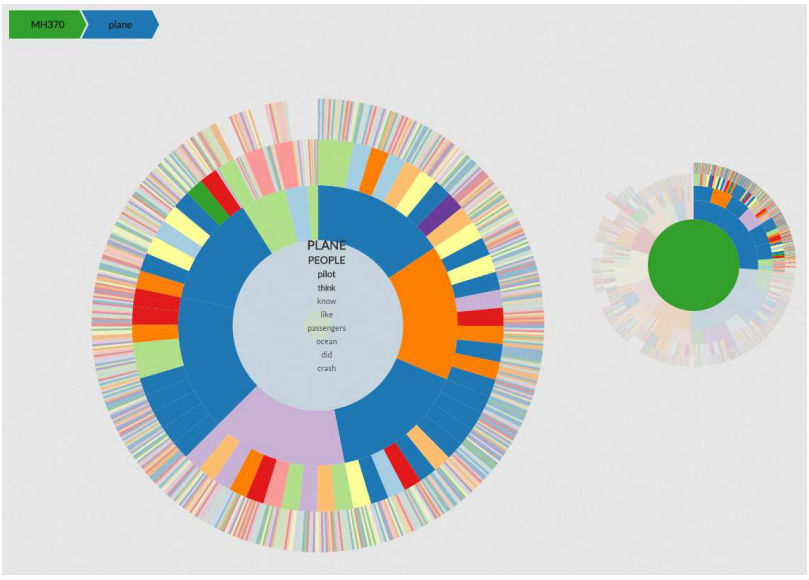

Рис. 15. Система *Hiérarchie*. Визуализация иерархии с помощью круговой диаграммы. С разрешения авторов [29].

до корня текущего уровня. Справа всегда находится «якорь», который отображает всю иерархию с выбранными темами и подтемами. Таким образом, пользователь всегда видит, на каком уровне иерархии он находится.

Рассмотрим взаимодействие пользователя с системой на примере визуализации твитов и новостей о пропавшем малазийском боинге MH-370 (рис. 15). Целью является анализ теорий, согласно которым пропал самолет. Была построена модель из 10 тем для каждого уровня. При выборе темы «*plane, people, pilot, think, know*», которая освещает различные теории, пользователь переходит на следующий уровень иерархии. Пользуясь навигацией по уровням, он находит наиболее обсуждаемые теории: самолет приземлился, самолет разбился, самолет был захвачен террористами, пилот разбил самолет в попытке суицида. Если выбрать тему о крушении, то на этом уровне также есть подтемы: самолет разбился в океане, на суше, из-за технических неполадок.

Достоинства: хорошая интерпретируемость, интуитивно понятный интерфейс; высокая детализация тем и подтем; нет ограничения для число уровней иерархии; удобная навигация по иерархии.

Недостатки: невозможность задать число тем для каждого уровня; число терминов темы также фиксировано и не может быть изменено пользователем; просмотр документов пока невозможен; медленная реакция системы на действия пользователя.

Ссылки:
`https://github.com/mlvl/Hierarchie` — сайт системы *Hiérarchie*;
`http://mlvl.github.io/Hierarchie` — пример визуализации.

Система TextFlow

Система *TextFlow* [11] предназначена для визуализации *динамических* (temporal) тематических моделей таких коллекций, в которых каждый документ имеет отметку времени создания или публикации. *TextFlow* позволяет находить и анализировать переломные события в темах — их появление и исчезновение, разделение и слияние.

Система состоит из трёх основных компонент (рис. 16, слева). *Препроцессор* извлекает основной текст документов. *Тематический анализатор* строит динамическую тематиче-

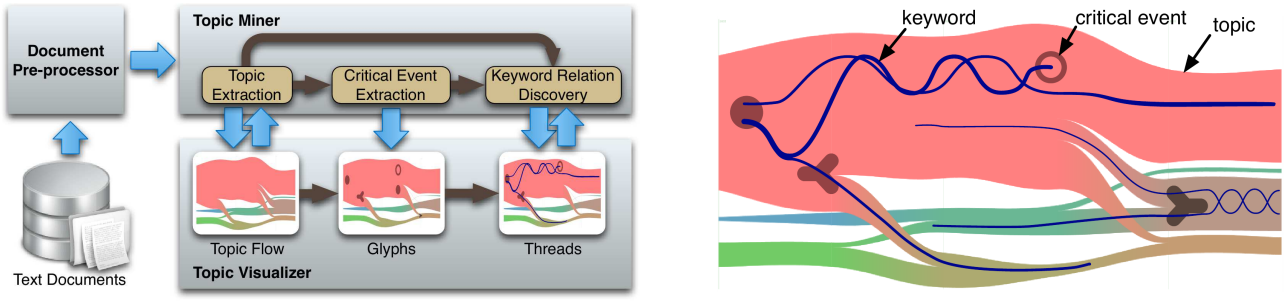

Рис. 16. Система *TextFlow*. Архитектура системы (слева) и пример визуализации (справа): 4 потока тем, 4 переломных изменения и 5 нитей ключевых слов. С разрешения авторов [11].

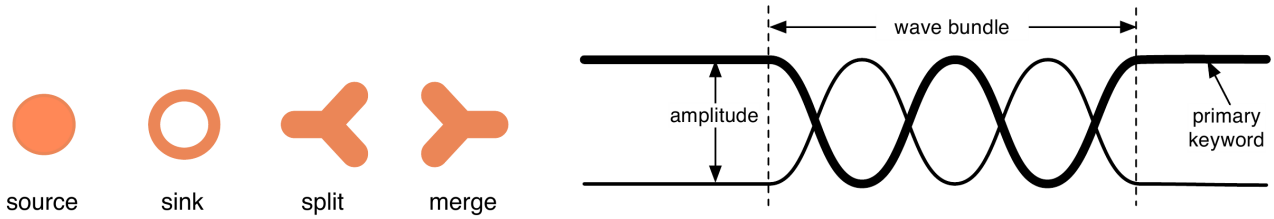

Рис. 17. Система *TextFlow*. Глифы (слева) отображают переломные моменты: исток, сток, разделение, слияние. Визуальные атрибуты (справа) отображают взаимодействия нитей: амплитуда и волновой пучок. С разрешения авторов [11].

скую модель, выявляет моменты слияния и разделения тем, а также переломные события и взаимосвязи между терминами тем. *Тематический визуализатор* отображает результаты в трёх компонентах визуализации.

Изменение тем как поток. В *TextFlow* изменения тем с течением времени изображаются в виде графа потоков (рис. 16, справа). Время откладывается вдоль горизонтальной оси. Изменяющаяся высота потока вдоль вертикальной оси пропорциональна количеству документов, которые относятся к данной теме в данный момент времени. Потоки могут разделяться и сливаться. При разделении (слиянии) главной веткой считается та, тематика которой наиболее близка к тематике потока до разделения (после слияния). При соединении потоков цвет получившейся ветви получается смешением цветов соответствующих потоков. Пропорции смешивания определяются высотами исходных веток. Аналогичный механизм применяется при разделении потоков.

Переломные изменения как глифы. Для отображения появления, исчезновения, соединения и разделения потоков были выбраны четыре глифа (рис. 17, слева). Они накладываются поверх изображения потоков в моменты событий, обнаруженных тематическим анализатором. Чем больше глиф, тем важнее обозначаемое им событие.

Корреляции ключевых слов как нити. Для изображения взаимосвязи определенного термина (ключевого слова) темы с другими её терминами используется визуальный примитив, называемый *нитью* (рис. 16). Моменты времени, когда появляется ключевое слово и когда оно исчезает, соединяются кривой линией с волновым эффектом (рис. 17, справа). Если несколько нитей взаимодействуют друг с другом, то для изображения совместного появления этих ключевых слов используется волновой пучок в том интервале времени, когда ключевые слова часто совместно появлялись в документах. Амплитуда пучка пропорциональна частоте появления всех ключевых слов темы: чем она больше, тем чаще появляются соответствующие слова в данный момент времени.

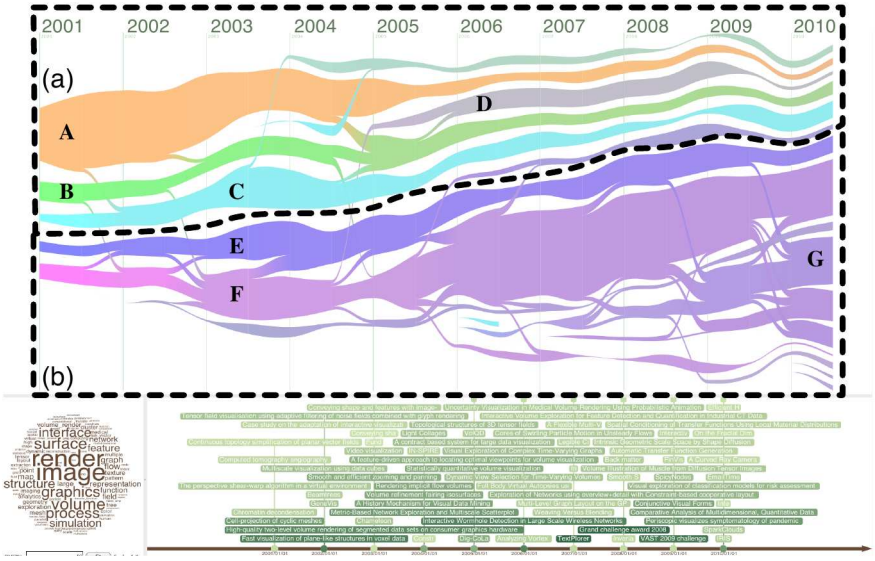

Рис. 18. Система *TextFlow* на примере визуализации коллекции статей научных конференций Vis и InfoVis. Основной графический интерфейс *TextFlow* и дополнительные компоненты: облако тегов (*слева внизу*) и временная шкала (*справа внизу*). С разрешения авторов [11].

Дополнительные средства: облако тегов и временная шкала. Описанные выше три компонента отображают только динамику изменений тем, но скрывают детальную текстовую информацию. Поэтому в *TextFlow* были добавлены два дополнительных компонента (рис. 18, внизу): облако тегов и временная шкала, которые при выделении темы отображают наиболее значимые термины и предложения этой темы.

В системе *TextFlow* есть две основные функции пользовательского интерфейса: наведение курсора на элемент и его выбор. При наведении пользователь видит агрегированную информацию, что облегчает ему выбор элемента. Например, при наведении на тему будут отображаться её наиболее существенные термины и ветки, исходящие из нее. При выборе пользователь видит более детальную информацию об элементе (например, облако тегов для темы). Последнее нужно для того, чтобы система могла рекомендовать пользователю шаги дальнейшей навигации по коллекции (например, схожие по составу темы или слова, которые часто появляются вместе с выбранным ключевым словом).

Рассмотрим взаимодействие пользователя с системой на примере визуализации коллекции статей научных конференций IEEE Visualization (Vis) and IEEE Information Visualization (InfoVis) с 2001 по 2010. На рис. 18 показан результат визуализации, где сразу видно несколько особенностей. Темы *A*–*D*, относящиеся к конференции Vis, сливались и разделялись до 2006, но затем развивались независимо друг от друга. Темы конференции InfoVis *E*–*G*, наоборот, активно взаимодействовали друг с другом на всём интервале времени. При этом тема *F* «исследование/аналитика» является «главным» потоком, порождающим большое количество ответвлений. При детальном анализе темы *F* (рис. 19) отчетливо видно, что в середине потока есть исток (a). При анализе относящихся к нему документов выяснилось, что большинство из статей относится к теме «аналитика». Это не случайное совпадение, т. к. в 2006 году впервые был проведен симпозиум IEEE VAST. Таким образом, при детальном или общем анализе визуализации пользователь может найти интересные особенности в коллекции.

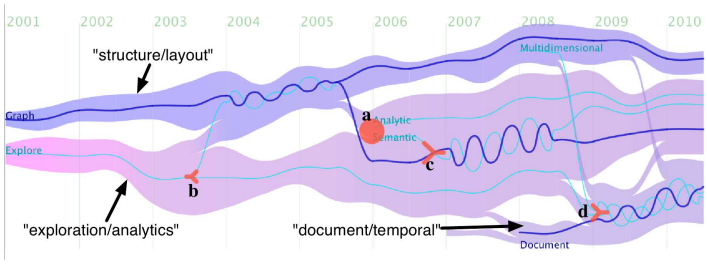

Рис. 19. Система *TextFlow* на примере визуализации коллекции статей научных конференций Vis и InfoVis. Детализация темы F конференции InfoVis. С разрешения авторов [11].

Достоинства: богатый и удобный интерфейс для исследования тем во времени; визуализация переломных изменений и ключевых слов.

Недостатки: непонятно, как выбирать главное ключевое слово для отображения нитей; во временной шкале отображаются только отрывки документов, причем не всех; неудобно сравнивать темы в интерфейсе; работает на коллекциям малого объёма.

Пример визуализации:
`http://cgcad.thss.tsinghua.edu.cn/shixia/publications/textflow/video.avi`

Система HierarchicalTopics

Система *HierarchicalTopics* [14] совмещает в себе построение и интерактивную визуализацию иерархических тематических моделей в их динамическом развитии (рис. 20).

Система реализует четыре стадии обработки исходных данных (рис. 21): (A) накопление данных, (B) предварительная обработка, параллельное тематическое моделирование и суммаризация текстовых документов, (C) построение иерархического дерева путём слияния тематических кластеров, (D) визуализация. Первые две стадии реализуются в режиме оффлайн, последние две — в режиме интерактивного взаимодействия пользователя с системой.

Визуализация *HierarchicalTopics* состоит из двух синхронизированных представлений: отображение иерархии *(Hierarchical Topic view)* и отображение потоков тем *(Hierarchical ThemeRiver)*.

Hierarchical Topic view. В этом представлении пользователь может активно взаимодействовать с системой (рис. 22). Помимо стандартного масштабирования и прокрутки, он может использовать лупу (для увеличения размера шрифта слов темы) и выделитель (для выделения слов темы). Также пользователь может изменять структуру иерархии: группировать несколько вершин дерева в одну, поглощать вершины и сворачивать их (рис. 23). Имеется возможность подписать каждую вершину группы тем.

Hierarchical ThemeRiver. Для отображения динамических изменений в группах тем используется модуль визуализации *ThemeRiver* [20]. Работа пользователя начинается с главной панели, где отображаются изменения самых верхних тем иерархии (рис. 24A). Высота каждой ленты (т. е. части графика, соответствующей конкретной теме) вычисляется как сумма высот листьев соответствующей вершины. При наведении курсора на ленту на панели предварительного просмотра отображаются изменения, произошедшие в дочерних узлах, рис. 24B). Для сравнения нескольких групп тем в *HierarchicalTopics* используется гибкая структура панелей. Чтобы сравнить различные группы, пользователь может выбрать нужную ему ленту на графике, для которой построится под-панель (рис. 24C), показывающая следующий уровень иерархии для выбранной темы.

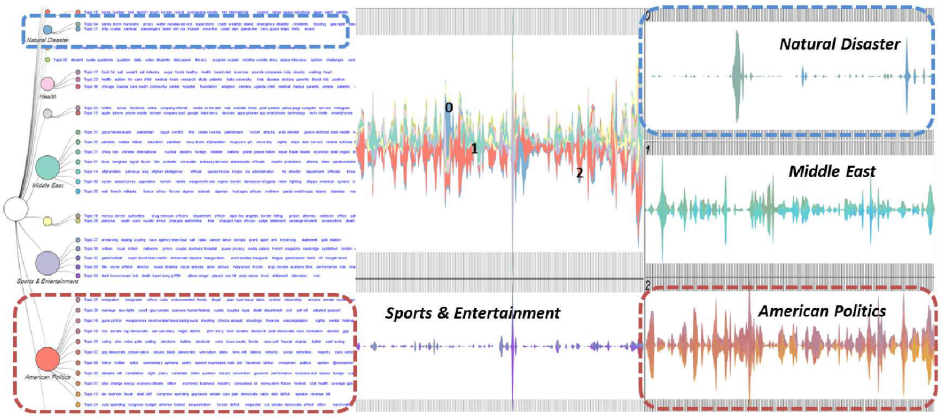

Рис. 20. Система *HierarchicalTopics*. Основные элементы пользовательского интерфейса. С разрешения авторов [14].

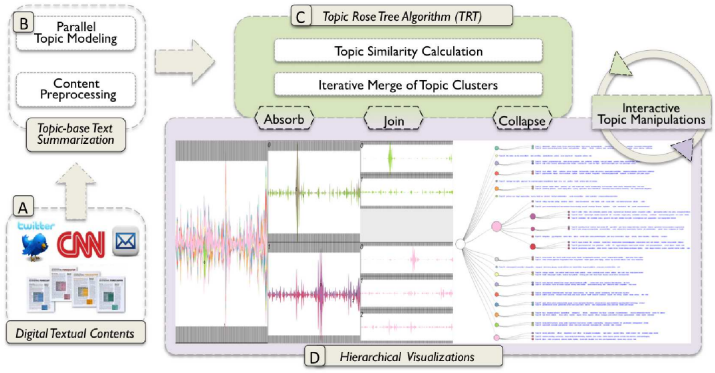

Рис. 21. Система *HierarchicalTopics*. Архитектура. С разрешения авторов [14].

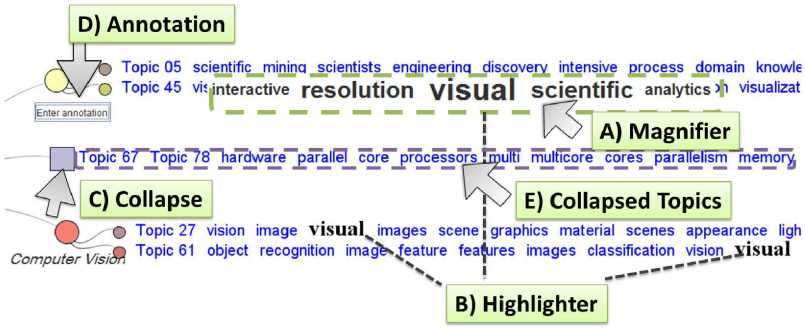

Рис. 22. Система *HierarchicalTopics*. Основные действия пользователя может совершать в *Hierarchical Topic view*: (A) лупа, (B) выделитель, (C) сворачивание вершин (при этом форма вершины становится прямоугольником), (D) аннотация, которую пользователь может добавить к любой вершине. С разрешения авторов [14].

Цветовая схема. Для плавного перехода между панелями используются 12 специально подобранных когерентных цветов, которые присваиваются лентам на графике. Ленты дочерних узлов окрашиваются в тот же оттенок, но другой яркости и насыщенности.

Детализация текстовых документов возможна после выделения темы. В представлении *Hierarchical ThemeRiver* пользователь может включить режим «временно́й поддерж-

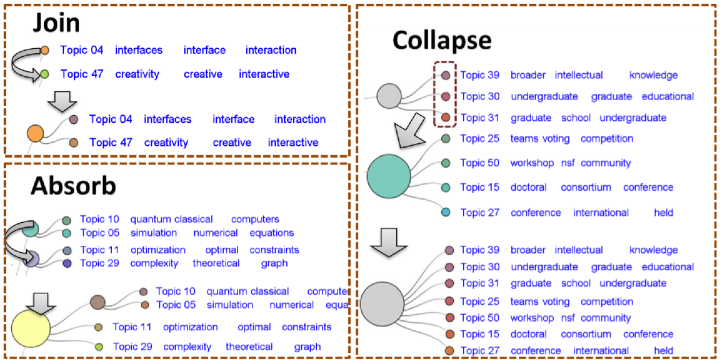

Рис. 23. Система *HierarchicalTopics*. Три операции, с помощью которых пользователь может изменить иерархию тем: соединение (join), поглощение (absorb) и свёртывание (collapse). С разрешения авторов [14].

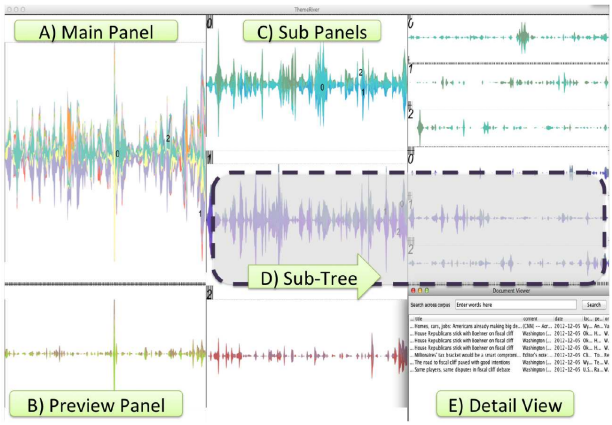

Рис. 24. Система *HierarchicalTopics*. Основной интерфейс *Hierarchical ThemeRiver*. С разрешения авторов [14].

ки» и просматривать множества документов, которые были опубликованы в какой-либо конкретный период времени (рис. 24E).

Достоинства: соединение в одном приложении визуализации как иерархической, так и динамической модели; интуитивно понятный интерфейс; возможность интерактивного просмотра и изменения иерархии тем; возможность детализации текстов документов за любой период времени.

Недостатки: определение высоты ленты на временном графике через сумму высот листьев не является полезной характеристикой для исследования коллекции; при большом количестве тем график перестает быть визуально понятным.

Ссылки:
`http://youtu.be/Vi1FP5kAbOU` — пример визуализации.

Система RoseRiver

Система *RoseRiver* [13] разработана на основе системы *TextFlow* и предназначена для анализа динамики изменений в иерархических тематических моделях. Она позволяет сливать и разделять темы, выбирать уровень детализации иерархии путём задания разреза дерева и прослеживать изменения выбранного множества тем во времени.

Сначала по коллекции документов строится последовательность т. н. эволюционных деревьев тем, которые представляют иерархию тем в коллекции в различные моменты

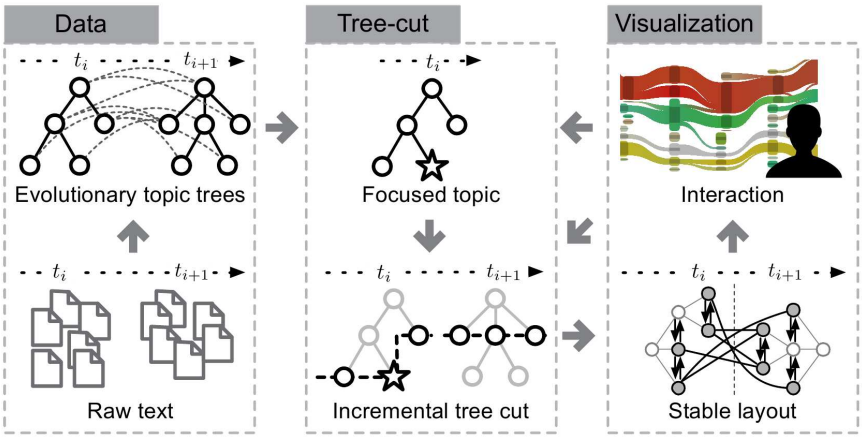

Рис. 25. Система *RoseRiver* состоит из трех компонентов: модуль обработки данных и тематического моделирования, модуль построения разрезов и модуль визуализации. С разрешения авторов [13].

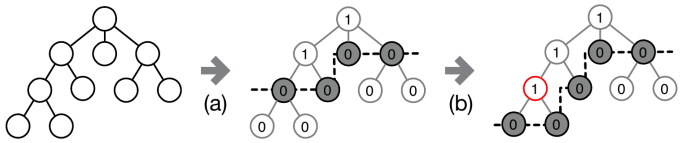

Рис. 26. Система *RoseRiver*. Создание разреза и его модификация: (a) выделены вершины разреза; (b) тема перестала быть фокусом, сгенерирован новый разрез. С разрешения авторов [13].

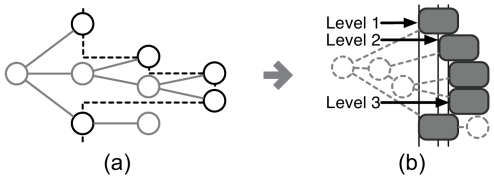

Рис. 27. Система *RoseRiver*. Вершины разреза (a) и их представление с небольшими сдвигами уровней вдоль оси времени (b). С разрешения авторов [13].

времени (рис. 25). Кроме последовательности деревьев на этом шаге также формируется множество пар схожих вершин-тем в деревьях, соответствующих смежным моментам времени. Система учитывает интересы пользователя, накапливая информацию о том, за изменением каких тем он следит. Для этого применяется техника построения деревьев с учётом степени интереса (*degree-of-interest, DOI*) [6].

В основе визуализации *RoseRiver* лежит *инкрементный алгоритм построения разреза в эволюционном дереве (incremental evolutionary tree cut algorithm). Разрезом дерева (tree cut)* называется такое множество вершин, что любой путь из корня дерева к его листу содержит только одну вершину из разреза (рис. 26). Идея алгоритма заключается в том, чтобы разрезы деревьев в последовательные моменты времени состояли из схожих вершин. Для этого пользователь фиксирует момент времени и выбирает одну или более основных вершин, называемых фокусными темами. На основе фокусных тем строится разрез дерева, называемый ключевым. По ключевому разрезу строится производное множество разрезов деревьев в смежные моменты времени, проходящих через вершины-

Рис. 28. Система *RoseRiver*. Четыре примера интерпретации потоков и полоски между ними: (a) возникновение темы; (b) тема мало изменяется; (c) тема сильно изменяется; (d) тема быстро возникает и исчезает. С разрешения авторов [13].

Рис. 29. Система *RoseRiver*. Расширение полоски темы с сохранением ее положения. С разрешения авторов [13].

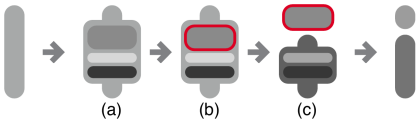

Рис. 30. Система *RoseRiver*. Пример разделения потоков тем. С разрешения авторов [13].

темы, схожие с темами ключевого разреза. Пользователь может исправить автоматически найденные разрезы в интерактивном режиме, чтобы улучшить интерпретируемость тем.

Визуализация потока тем в *RoseRiver* полностью основана на *TextFlow* (см. выше), поэтому далее будут описаны только дополнительные элементы, появившиеся в *RoseRiver*.

Вершины разреза. Каждая вершина из разреза представляется прямоугольником со скруглёнными углами, при этом уровни иерархии представляются небольшими сдвигами вправо вдоль оси времени (рис. 27).

Документы. Цветная полоска между двумя или более потоками тем, разделенными вершиной из разреза, обозначает число пар документов, относящихся к этим темам (рис. 28). Темная часть этой полоски обозначает документы, которые принадлежат обеим темам слева и справа от полоски (т. е. двум соседним деревьям). Высота темной части пропорциональна доле таких документов.

Цветовая схема. Фокусные темы имеют полностью насыщенный уникальный цвет (рис. 31, 32), в то время как производные темы (например, полученные в результате слияния) отличаются оттенками, т. е. сохраняется цветовая стратегия *TextFlow*. При этом цвет постепенно сводится к серому, если тема перестает быть похожей на темы-фокусы и уже не удовлетворяет интересам пользователя.

Детали. При наведении курсора на вершину-тему из разреза полоска внутри нее расширяется, показывая характерные для темы слова внутри расширенной области (рис. 29). Чтобы сохранить информацию о глубине такой вершины, расширяется только средняя часть полоски, концы при этом остаются неизменными.

Взаимодействие с системой. Пользователь может изменять потоки тем, сливая или разделяя темы. Для этого при выборе полоски вершины-темы она автоматически распадается на «под-полоски», которые обозначают темы, содержащиеся в потоке слева или справа. При выборе нескольких таких «под-полосок» можно слить воедино или разъединить несколько потоков тем (рис. 30).

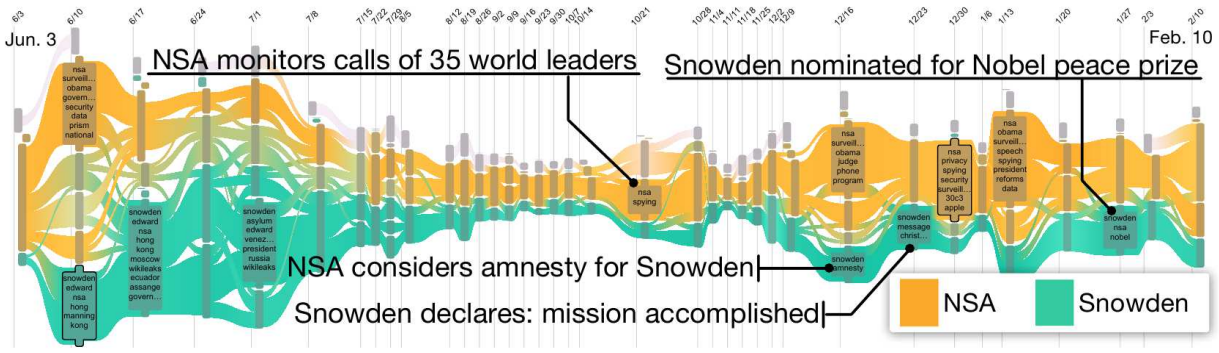

Рис. 31. Система *RoseRiver*. Коллекция новостей и твитов о *Prism* с июля 2013 по февраль 2014. С разрешения авторов [13].

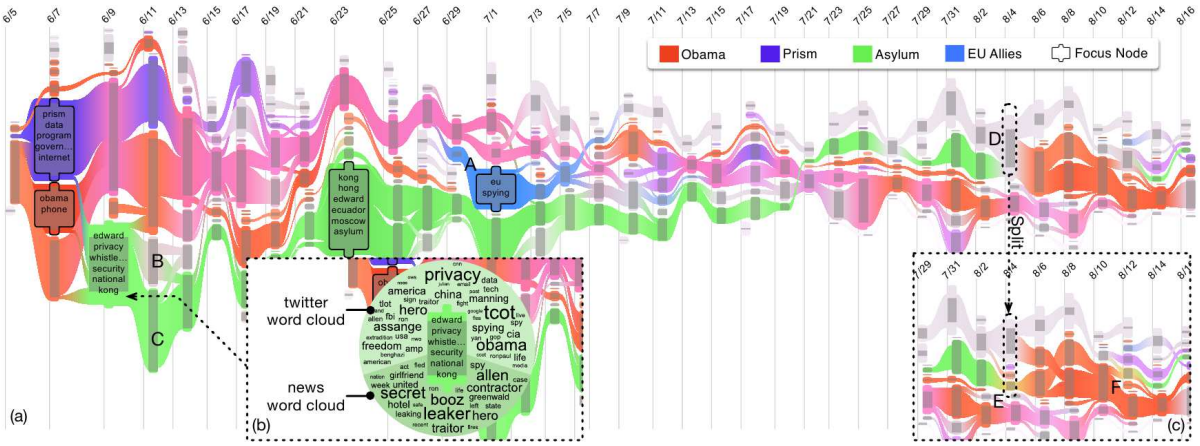

Рис. 32. Система *RoseRiver*. (a) Новости о PRISM с июня по август 2013. (b) Сравнение ключевых слов из Твиттера и новостей, длина дуг в облаке соответствует числу сообщений и статей. (c) После разделения двух потоков сгенерировано новое представление. С разрешения авторов [13].

Также пользователь может изменять фокусные темы в процессе работы. Это возможно с помощью простого выделения интересующей темы или с помощью поиска, если пользователь не видит нужную ему тему на экране.

Рассмотрим работу пользователя с системой на примере визуализации коллекции новостей и твитов о *Prism* (рис. 31). Для более детального рассмотрения были выбраны тема «*NSA*» (оранжевая) и тема «*Snowden*» (голубая). Как видно из рис. 31, голубая тема была намного популярнее в течение первого месяца, чем оранжевая тема, но с течением времени она стремительно теряла популярность, в то время как оранжевая тема почти не изменялась. Из этого можно сделать вывод, что оранжевая тема является перманентной, а голубая — событийной. То же самое можно сказать и о синей теме «*EU Allies*» (рис. 32.), которая появилась на короткое время и исчезла. Для выявления причин такого изменения темы пользователь может разделить ее на подтемы и проанализировать их. Например, на рис. 32D видно, что причиной угасания зеленой темы стало появление желтой темы.

Достоинства: комбинированная визуализация иерархической и динамической модели при сохранении всех преимуществ *TextFlow*; возможность интерактивного взаимодействия с системой для изменения отображаемой иерархии; облако тегов отображается не внизу интерфейса, а прямо рядом с контур-вершиной.

Недостатки: сложный интерфейс для неподготовленного пользователя; непонятно, как выбирать фокусные темы, когда пользователь заранее не знает, о чем коллекция; возможны сбои при построении эволюционных деревьев и контуров в них; нет навигации по конкретным деревьям в конкретный момент времени.

Ссылки:
`http://research.microsoft.com/en-us/um/people/weiweicu/RoseRiver` — сайт проекта;
`http://cgcad.thss.tsinghua.edu.cn/shixia/publications/RoseRiver/video.mp4` — пример визуализации.

Краткий обзор других систем визуализации

Система *MetaToMATo*(Metadata and TopicModel Analysis Toolkit) [30] предназначена для визуализации не только тематической модели, но и метаданных коллекции (рис. 33). *MetaToMATo* позволяет пользователь переименовать темы, удалять темы, добавлять различные типы метаданных — авторов, метку времени, географическую метку, метку источника и т. д. Также с помощью *MetaToMATo* можно осуществлять тематический поиск по метаданным.

Достоинства: удобный и функциональный интерфейс для отображения метаданных и контроля степени их влияния на отображаемые документы при поиске.

Недостатки: невозможность отображения схожих тем, невозможность улучшить модель (изменить слова темы, слить или разделить темы), список терминов темы состоит из самых вероятных слов, а не из наиболее характерных слов темы.

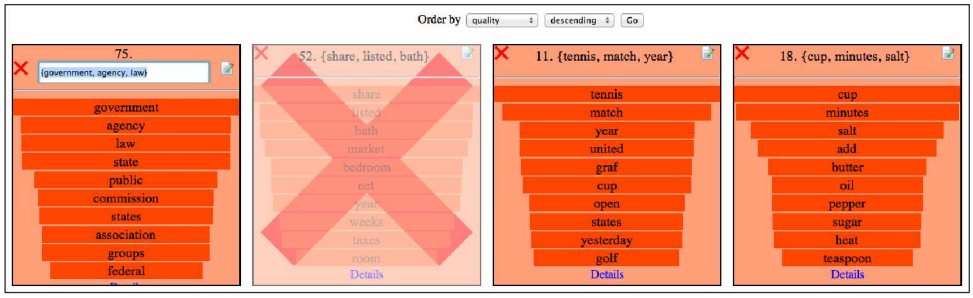

Рис. 33. Система *MetaToMATo*. Страницы темы. С разрешения авторов [30].

Система *LDAvis* показывает, насколько темы весомы в коллекции, насколько они близки друг к другу, и из каких значимых терминов они состоят [28]. Для этого в *LDAvis* имеются два представления (рис. 34). Первое отображает темы в виде вершин, размер которых пропорционален $p(t)$, и расстояние между которыми пропорционально мере сходства тем. Второе показывает распределения $p(w\,|\,t)$ и $p(w)$ для релевантных терминов темы. Релевантность термина w относительно темы t определяется как

$$relevance(w, t\,|\,\lambda) = \lambda \log p(t\,|\,w) + (1-\lambda)\log\frac{p(t\,|\,w)}{p(w)},$$

где параметр λ $(0 \leqslant \lambda \leqslant 1)$ регулируется пользователем (авторы рекомендуют $\lambda = 0.6$). Также *LDAvis* группирует схожие темы в кластеры и обозначает каждый кластер уникальным цветом (максимальное число кластеров равно 10).

Достоинства: удобный интерактивный интерфейс для исследования распределений слов по темам и по коллекции, где степень релевантности терминов контролируется пользователем.

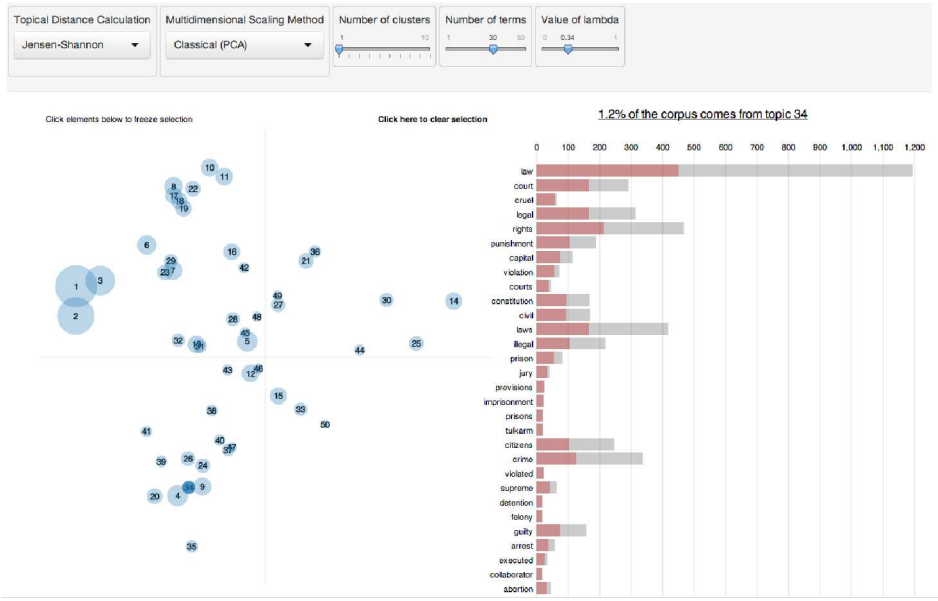

Рис. 34. Система *LDAvis*. С разрешения авторов [28].

Недостатки: невозможность просмотра документов и распределений тем по документам, ограниченное количество кластеров и отображаемых терминов, невозможность задать количество тем, отсутствие именования тем.

Сайт: `https://github.com/cpsievert/LDAvis`.

Пример: `http://cpsievert.github.io/LDAvis/reviews/vis`.

Система *TIARA (Text Insigt via Automated Responsive Analytics)* предназначена для визуализации динамических моделей, которая представляет темы как потоки в потоковом графе [31]. На каждом потоке в каждый момент времени изображается облако слов темы, популярных в данный момент (рис. 35). Высота потока пропорциональна количеству документов, относящихся к теме. В *TIARA* пользователь может улучшать тематическую модель, соединяя/разъединяя потоки, а также осуществлять тематический поиск по коллекции. Продолжением этой системы стал *TextFlow*.

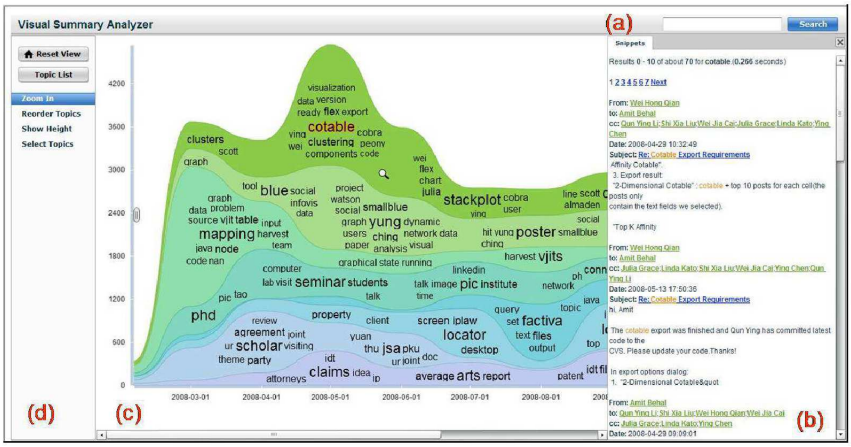

Рис. 35. Система *TIARA*. С разрешения авторов [31].

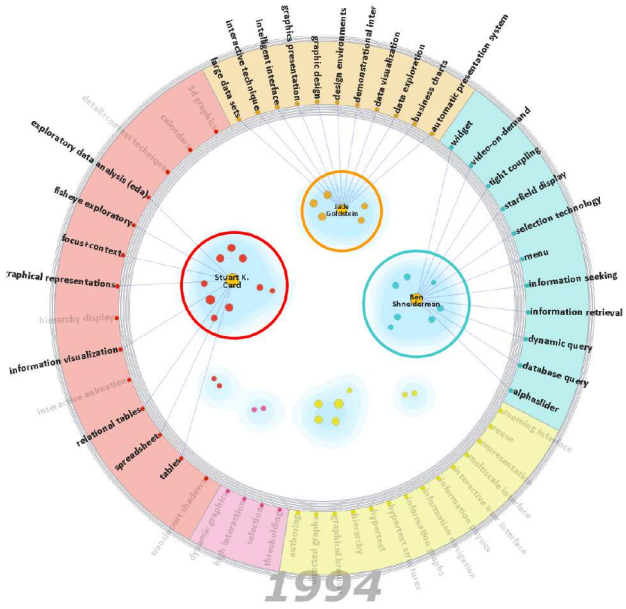

Рис. 36. Система *SolarMap*. С разрешения авторов [4].

Система *SolarMap* предназначена для визуализации динамических моделей (рис. 36) [4]. В *SolarMap* документ разбивается на аспекты (*facets*). Аспектами могут быть даты или структурные части документов. Например, если коллекция — это статьи с описанием различных заболеваний, то аспектами могут быть разделы «симптомы», «лечение», «профилактика». Кроме того, из текста извлекаются именованные сущности (*named entity*). Аспекты изображаются в виде тонких колец, внутри них группируются сущности в виде облака, вокруг которых изображаются ключевые слова в виде толстого кольца, разбитого на части, относящиеся к группам сущностей. В качестве аспектов можно задавать даты, и, меняя тонкие кольца, следить за развитием тем в коллекции.

Достоинства: визуализация различных метаданных, включая метки времени, возможность отслеживания конкретной группы сущностей с течением времени.

Недостатки: отсутствие просмотра документов, отсутствие именования тем, невозможность улучшения модели.

Пример: `http://nancao.org/projects/solarmap.tml`.

Система *TextWheel* предназначена для визуализации новостных потоков [12]. Она состоит из трех компонентов: «конвейера» документов, «колеса» ключевых слов и временной шкалы (рис. 37). Временная шкала отображает значимость документов с течением времени. U-образный конвейер содержит глифы, обозначающие документы, которые соединяется с колёсами ключевых слов. При этом слова соединяются с документами через два узла, обозначающих положительное и отрицательное отношение.

Достоинства: пользователь может контролировать скорость конвейера, соединять глифы, детализировать визуализацию и осуществлять повторную кластеризацию.

Недостатки: нет отображения тем, нет поиска документов, отсутствуют средства визуализации всей коллекции.

Система *ThemeDelta* предназначена для визуализации динамических моделей [17]. *ThemeDelta* извлекает из коллекции т.н. *тренды* (ключевые слова) и группирует их в темы в каждый момент времени (рис. 38). Тренды изображаются в виде волнистых линий, цвет которых зависит от темы, толщина — от значимости слова в теме. Моменты времени

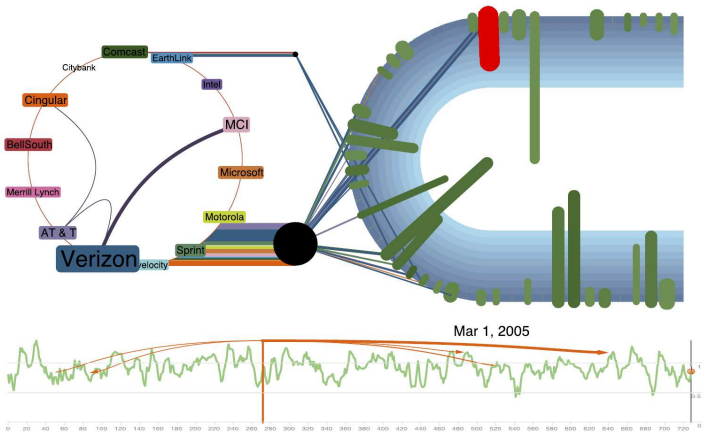

Рис. 37. *TextWheel.* С разрешения авторов [12].

изображаются в виде вертикальных осей, на которых расположены группы трендов (темы). Пользователь может осуществлять поиск трендов, сортировать и фильтровать их, детализировать визуализацию нескольких трендов и трендов, тесно связанных с ними.

Достоинства: возможность отследить изменения темы с течением времени.

Недостатки: избыточная детальность, отсутствие визуализации распределений слов по темам, отсутствие именования тем, невозможность улучшения тематической модели.

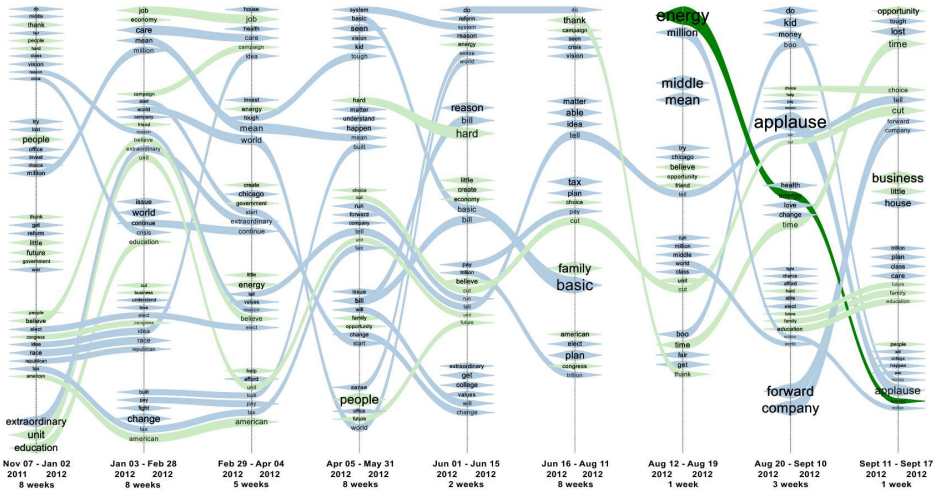

Рис. 38. Система *ThemeDelta.* С разрешения авторов [17].

Система *D-VITA* предназначена для интерактивной визуализации динамических моделей [19]. Пользовательский интерфейс системы состоит из следующих компонентов (рис. 39): потоковый граф для отображения тем (толщина потока пропорциональна релевантности темы); потоковый граф для отображения слов темы; круговая диаграмма для отображения пропорций тем в документе; модуль просмотра документов с возможностью поиска схожих по тематике документов.

Достоинства: удобный интерактивный интерфейс для навигации по коллекции, возможность осуществлять тематические запросы.

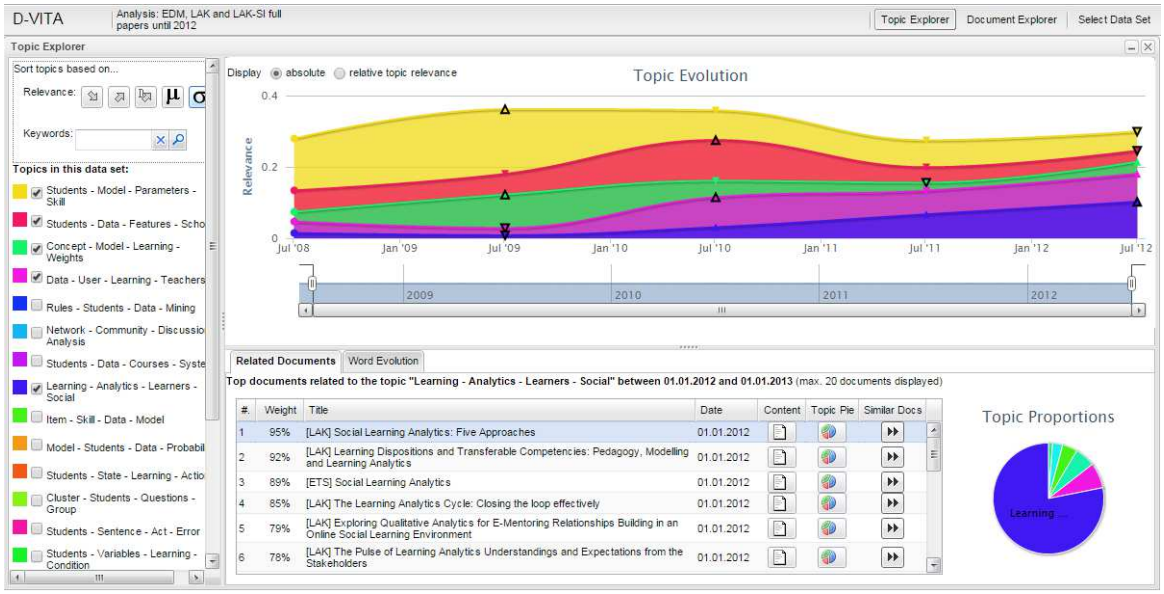

Рис. 39. Система *D-VITA*. С разрешения авторов [19].

Недостатки: отсутствие именования тем, отсутствие отображения схожих тем, невозможность улучшения тематической модели, в частности, отсутствие настройки степени сглаживания тем во времени.

Сайт: `https://github.com/rwth-acis/D-VITA`.

Пример: `http://monet.informatik.rwth-aachen.de/DVita`.

Система *TopicPanorama* предназначена для визуализации иерархических моделей [25]. Пользовательский интерфейс состоит из трех компонентов: основной панели для визуализации, информационной панели и панели управления. *TopicPanorama* представляет иерархическую структуру тем в виде дерева (рис. 40), при этом корреляции между темами отображаются с помощью областей разной цветовой насыщенности. Информационная панель показывает дополнительную информацию о темах и о документах, относящихся к теме. Панель управления позволяет пользователю задавать параметры визуализации, осуществлять поиск и изменять иерархическую структуру модели.

Достоинства: возможность отображения иерархической структуры в двух представлениях — в виде графа и в виде столбчатой диаграммы, окружающей граф.

Недостатки: отсутствие отображения распределений тем по документам, невозможность изменить иерархическую структуру и вручную именовать темы.

Сайт: `http://cgcad.thss.tsinghua.edu.cn/shixia/publications/TopicPanorama/index.htm`.

Пример: `http://cgcad.thss.tsinghua.edu.cn/shixia/publications/TopicPanorama/video_eng.mp4`

Графические библиотеки

Для визуализации тематических моделей в веб-интерфейсах используются также графические библиотеки общего назначения, наиболее известные — *Gephi* и *D3.js*.

Система *Gephi* — это платформа с открытым кодом для анализа и визуализации больших сетей. Её применение в тематическом моделировании основано на том, что разреженные матрицы распределений терминов в темах и тем в документах можно рассматривать как матрицы смежности для двудольных графов с вершинами двух типов — соответственно, терминов и тем, либо тем и документов.

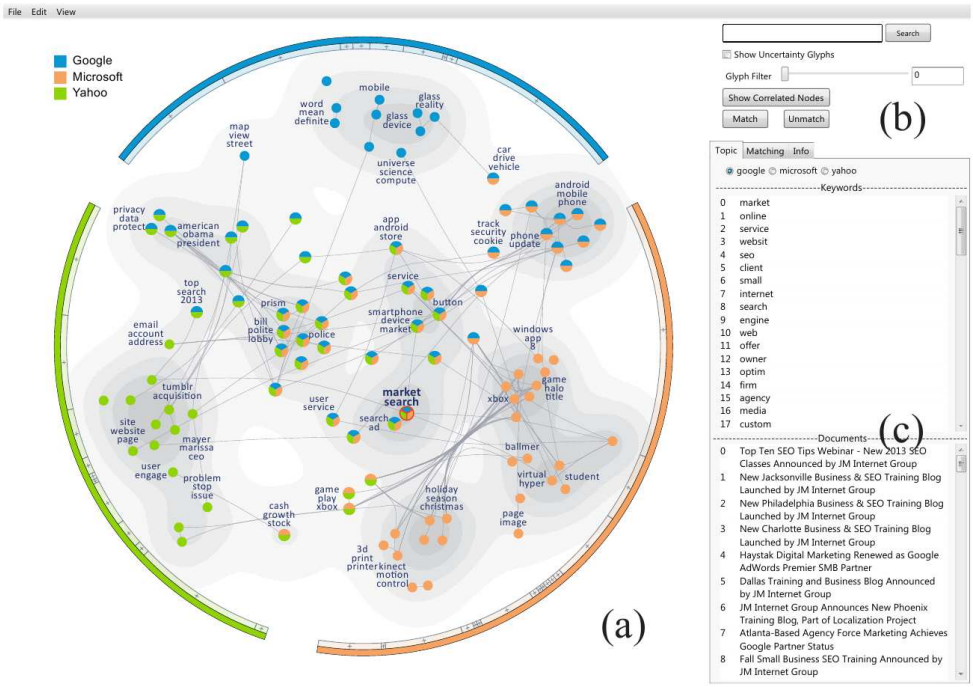

Рис. 40. Система *TopicPanorama*. С разрешения авторов [25].

Например, в [9] анализировалась коллекция текстов с обсуждениями медицинских препаратов для поддержания сна с сайта *www.patientslikeme.com*. Тематическая модель была построена с помощью пакета *MALLET*[3]. Исходные документы и полученные темы визуализировались с помощью системы *Gephi*. Документы отображаются точками, темы — своими названиями. Для автоматического размещения вершин графа в *Gephi* используются методы многомерного шкалирования (рис. 41).

На рис. 42 показаны два варианта интерактивной визуализации двудольного графа, в которых темы отображаются более крупными вершинами, термины — более мелкими.

Сайт: `http://gephi.github.io`.

Пример: `http://dig-eh.org/topic-modeling-and-gephi-a-work-in-progress`.

Библиотека *D3.js* (Data-Driven Documents) — это JavaScript-библиотека с открытым кодом для создания и контроля динамических и интерактивных графических элементов, которые могут отображаться в веб-браузере.

D3.js предоставляет мощный интерфейс для построения интерактивных визуализаций на основе больших данных и имеет широкое применение в различных областях, в том числе в тематическом моделировании. Например, рассмотренные выше системы *LDAvis* и *Hiérarchie* используют *D3.js* для визуализации тематических моделей.

Сайт: `http://d3js.org`.

Пример: `https://github.com/mbostock/d3/wiki/Gallery#visual-index`.

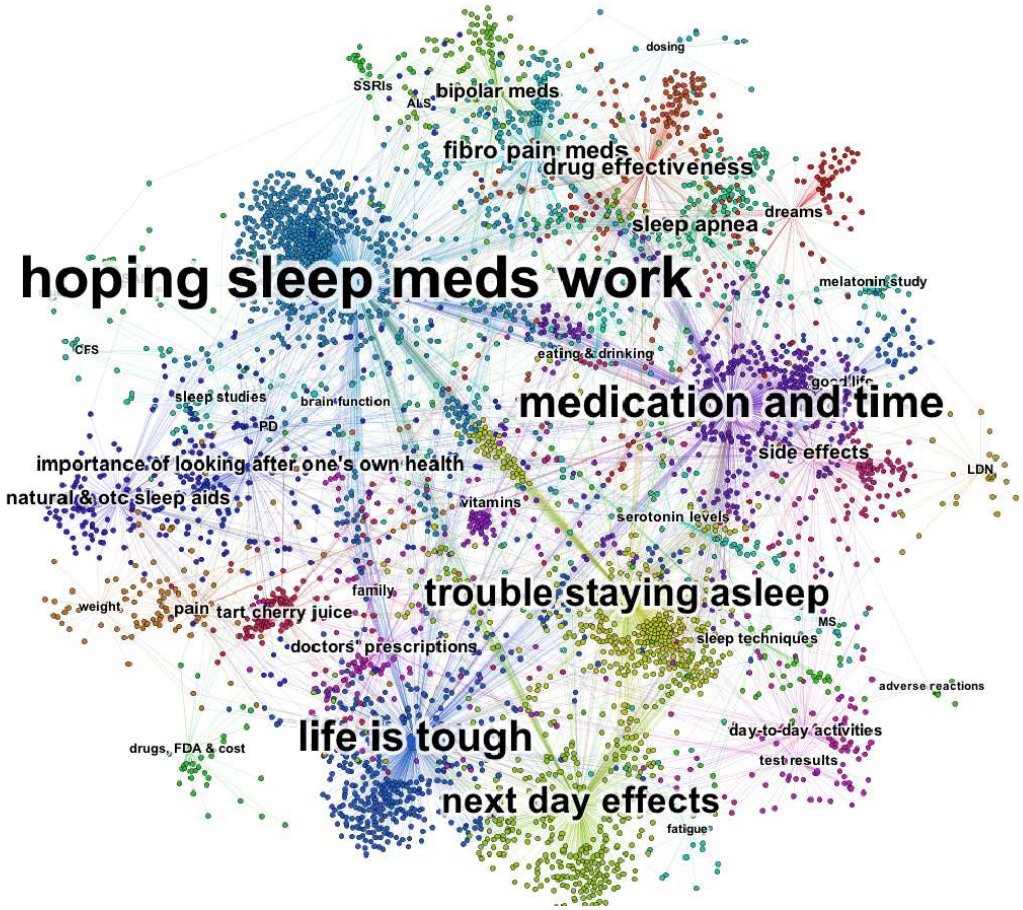

Рис. 41. Визуализация тематической модели LDA, построенной *MALLET*, с помощью *Gephi*. С разрешения авторов [9].

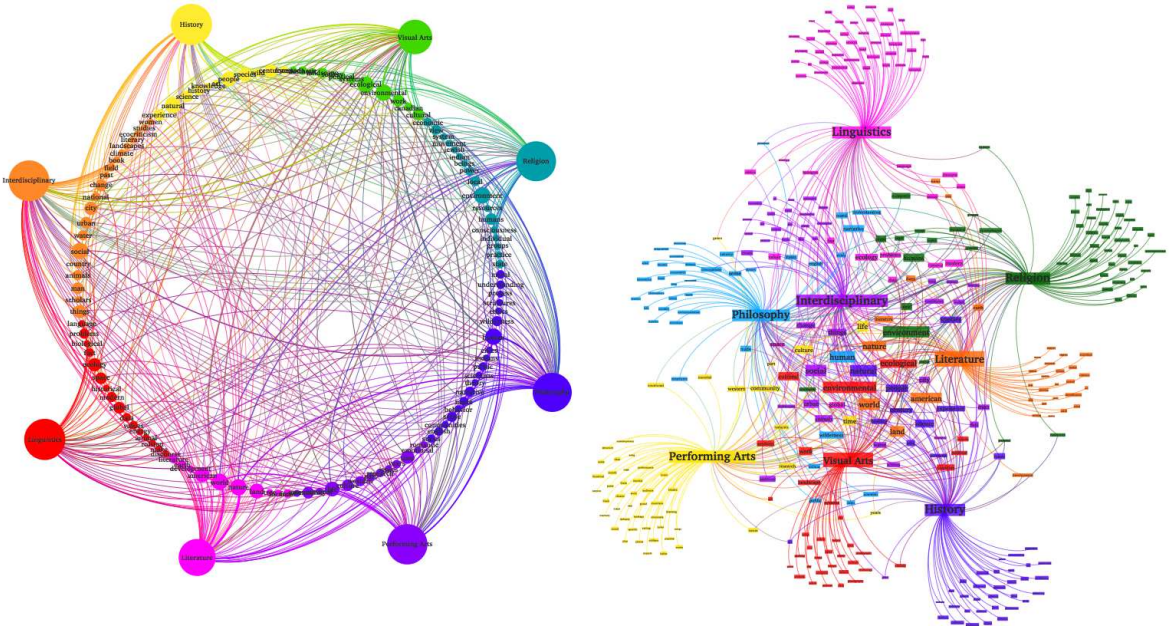

Рис. 42. Визуализация с помощью *Gephi*.

Таблица 1. Средства визуализации тематических моделей.

	TMVE	Termite	TopicNets	Hierarchie	iVisClustering	TextFlow	HierarchicalTopics	RoseRiver	MetaToMATO	LDAvis	TIARA	SolarMap	TextWheel	ThemeDelta	D-VITA	TopicPanorama
список тем	+	+	−	−	+	−	+	−	+	+	+	−	−	−	+	+
отображение схожих тем	+	−	+	−	+	+	−	+	+	+	+	+	−	+	−	+
именование тем:																
• автоматическое	+	+	+	+	+	+	+	+	+	−	−	−	−	−	+	+
• ручное	−	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−
отображение документов	+	+	+	−	+	+	+	+	+	−	+	−	+	?	+	+
• схожих документов	+	+	+	−	+	−	−	−	+	−	+	−	−	−	+	+
распределения:																
• $p(t \mid d)$	+	−	+	−	+	?	−	?	+	−	?	−	−	−	+	−
• $p(w \mid t)$	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
• $p(t)$	+	−	+	+	−	+	−	+	+	+	+	−	−	−	+	−
модальности:																
• время	−	−	+	−	−	+	+	+	+	−	+	+	+	+	+	−
• иерархии	−	−	+	+	+	−	+	+	−	−	−	+	−	−	−	+
• авторы	−	−	+	−	−	−	−	−	+	−	−	+	+	−	−	−
• другие	−	−	+	−	−	−	−	−	+	−	−	+	+	−	−	−
детализация	−	−	+	+	+	+	+	+	−	−	+	?	+	+	+	+
поисковые запросы	+	−	+	−	−	−	−	+	+	−	+	−	−	+	+	+
изменение модели:																
• повторное построение	−	−	−	−	+	+	−	+	−	−	+	−	+	−	−	+
• слияние или разделение тем	−	−	−	−	+	+	−	+	−	−	+	−	−	−	−	−
• удаление тем	−	−	−	−	+	+	−	+	+	−	−	−	−	−	−	−
• изменение веса терминов	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−
открытый код	+	+	−	+	−	−	−	−	−	+	−	−	−	−	+	+

Заключение

Важной тенденцией современных средств визуализации тематических моделей является использование веб-интерфейсов. Исследователи и разработчики стремятся возложить на мощные сервера всю подготовительную и вычислительную работу, связанную с накоплением коллекции, её предварительной обработкой, построением модели и поисковых индексов, генерацией визуальных представлений. Пользователям таких систем остаётся управлять параметрами визуального представления и средств навигации. Некоторые системы предполагают также обратную связь с пользователем, возможность внесения экспертных оценок или исправлений с целью улучшения модели.

В настоящее время не сложилось единого понимания, каким должен быть идеальный пользовательский интерфейс для тематического поиска и навигации по большим коллекциям текстовых документов. Отсюда большое разнообразие идей и систем для визуализации. Многие из них являются исследовательскими и далеки от стадии коммерческого использования. Некоторые из них имеют открытый исходный код.

[3] *MALLET* (***MA****chine* ***L****earning for* ***L****anguag****E*** ***T****oolkit*) — пакет под Java для обработки естественного языка, классификации документов, тематического моделирования и других приложений машинного обучения к анализу текстов [26].

Не претендуя на полноту, подытожим данный обзор таблицей 1, в которой сделана попытка систематизации систем визуализации тематических моделей в разрезе основных функциональных требований.

Автор выражает признательность К. В. Воронцову за постановку задачи и внимание к работе.

Литература

[1] *Asuncion A., Welling M., Smyth P., Teh Y.* 2009. On smoothing and inference for topic models // *25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 27–34.

[2] *Blei D., Ng Y., Jordan I.* 2003. Latent Dirichlet allocation // *Journal of Machine. Learning Research*. 3: 993–1022.

[3] *Blei D.* 2012. Probabilistic topic models // *Communications of the ACM*. 77–84.

[4] *Cao N., Gotz D., Sun J., Lin Y., Qu H.* 2011. SolarMap: Multifaceted Visual Analytics for Topic Exploration // *Data Mining (ICDM), IEEE 11th International Conference*. 101–110.

[5] *Cano A. E., He Y., Xu R.* 2014. Automatic labelling of topic models learned from Twitter by summarisation // *52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, USA. ACL. 618–624.

[6] *Card S., Nation D.* 2002 Degree-of-interest trees: A component of an attention-reactive user interface // *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM. 231–245.

[7] *Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Blei D. M.* 2009. Reading tea leaves: How humans interpret topic models // *Neural Information Processing Systems (NIPS)*. 288–296.

[8] *Chaney A., Blei D.* 2012. Visualizing Topic Models // *Frontiers of computer science in China*. 55(4): 77–84.

[9] *Chen A., Eichler G.* 2013. Topic Modeling and Network Visualization to Explore Patient Experiences // *Visual Analytics in Healthcare Workshop*.

[10] *Chuang J., Manning C., Heer J.* 2012. Termite: Visualization Techniques for Assessing Textual Topic Models // *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM. 74–77.

[11] *Cui W., Liu S., Tan L., Shi C., Song Y., Gao Z, Tong X., Qu H.* 2011. TextFlow: Towards Better Understanding of Evolving Topics in Text // *Visualization and Computer Graphics, IEEE Transactions*. 17(12): 2412–2421.

[12] *Cui W., Qu H., Zhou H., Zhang W., Wolfe T., Skiena S.* 2012. Watch the story unfold with TextWheel: Visualization of large-scale news streams // *Transactions on Intelligent Systems and Technology (TIST)*. ACM. 3(2): 20.

[13] *Cui W., Liu S., Wu Z., Wei H.* 2014. How Hierarchical Topics Evolve in Large Text Corpora // *IEEE Transactions on Visualization and Computer Graphics*. 20(12): 2281–2290.

[14] *Dou W., Yu L., Wang X., Ma Z., Ribarsky W.* 2013. HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies // *Visualization and Computer Graphics, IEEE Transactions*. 19(12): 2002–2011.

[15] *Daud A., Li J., Zhou L., Muhammad F.* 2010. Knowledge discovery through directed probabilistic topic models: a survey // *Frontiers of computer science in China*. 4(2): 280–301.

[16] *Eades P.* 2010. A heuristic for graph drawing // *Congressus numerantium*. 42: 146–160.

[17] *Gad S., Javed W., Ghani S., Elmqvist N., Ewing T., Hampton N., Ramakrishnan N.* 2015. ThemeDelta: Dynamic Segmentations over Temporal Topic Models // *Visualization and Computer Graphics, IEEE Transactions*. 21(5): 672–685.

[18] *Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P.* 2012. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling // *ACM Transactions on Intelligent Systems and Technology (TIST).* 3(2):23.

[19] *Günnemann N., Jarke M.* 2013. D-VITA: A Visual Interactive Text Analysis System Using Dynamic Topic Mining // *BTW workshops.* 237–246.

[20] *Havre S., Hetzler B., Nowell L.* 2002. ThemeRiver: visualizing thematic changes in large document collections // *Visualization and Computer Graphics, IEEE Transactions.* 17(12):9–20.

[21] *Hofmann T.* 1999. Probabilistic latent semantic indexing // *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM. 50–57.

[22] *Kuhn W.* 1995. The Hungarian method for the assignment problem // *Naval Research Logistic Quarterly.* 2:83–97.

[23] *Lau J. H., Grieser K., Newman D., Baldwin T.* 2011. Automatic labelling of topic models // *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Stroudsburg, PA, USA. ACL. 1536–1545.

[24] *Lee H., Kihm J., Choo J., Stasko J., Park H.* 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling // *Computer Graphics Forum.* 31(3):1155–1164.

[25] *Liu S., Wang X., Chen J., Zhu J., Guo B.* 2014. TopicPanorama: a Full Picture of Relevant Topics // *Visual Analytics Science and Technology (VAST).* 183–192.

[26] *McCallum A. K.* MALLET: A Machine Learning for Language Toolkit, 2002. `http://mallet.cs.umass.edu`.

[27] *Mei Q., Shen X., Zhai C.* 2007. Automatic labeling of multinomial topic models // *13th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, New York, NY, USA. 490–499.

[28] *Sievert C., Shirley K.* 2014. LDAvis: A method for visualizing and interpreting topics // *Workshop on Interactive Language Learning, Visualization, and Interfaces.*

[29] *Smith A., Hawes T., Myers M.* 2014. Hiérarchie: Interactive Visualization for Hierarchical Topic Models // *Sponsor: Idibon.* 71.

[30] *Snyder J., Knowles R., Dredze M., Gormley M., Wolfe T.* 2013. Topic Models and Metadata for Visualizing Text Corpora // *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* 10:5.

[31] *Wei F., Liu S., Song Y., Pan S., Zhou M., Qian W., Shi L., Tan L., Zhang Q.* 2010. TIARA: A Visual Exploratory Text Analytic System // *16th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM. 153–162.

References

[1] *Asuncion A., Welling M., Smyth P., Teh Y.* 2009. On smoothing and inference for topic models. *25th Conference on Uncertainty in Artificial Intelligence.* AUAI Press. 27–34.

[2] *Blei D., Ng Y., Jordan I.* 2003. Latent Dirichlet allocation. *Journal of Machine. Learning Research.* 3:993–1022.

[3] *Blei D.* 2012. Probabilistic topic models. *Communications of the ACM.* 77–84.

[4] *Cao N., Gotz D., Sun J., Lin Y., Qu H.* 2011. SolarMap: Multifaceted Visual Analytics for Topic Exploration. *Data Mining (ICDM), IEEE 11th International Conference.* 101–110.

[5] *Cano A. E., He Y., Xu R.* 2014. Automatic labelling of topic models learned from Twitter by summarisation. *52nd Annual Meeting of the Association for Computational Linguistics.* Baltimore, MD, USA. ACL. 618–624.

[6] *Card S., Nation D.* 2002 Degree-of-interest trees: A component of an attention-reactive user interface. *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM. 231–245.

[7] *Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Blei D. M.* 2009. Reading tea leaves: How humans interpret topic models. *Neural Information Processing Systems (NIPS)*. 288–296.

[8] *Chaney A., Blei D.* 2012. Visualizing Topic Models. *Frontiers of computer science in China*. 55(4): 77–84.

[9] *Chen A., Eichler G.* 2013. Topic Modeling and Network Visualization to Explore Patient Experiences. *Visual Analytics in Healthcare Workshop*.

[10] *Chuang J., Manning C., Heer J.* 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM. 74–77.

[11] *Cui W., Liu S., Tan L., Shi C., Song Y., Gao Z, Tong X., Qu H.* 2011. TextFlow: Towards Better Understanding of Evolving Topics in Text. *Visualization and Computer Graphics, IEEE Transactions*. 17(12): 2412–2421.

[12] *Cui W., Qu H., Zhou H., Zhang W., Wolfe T., Skiena S.* 2012. Watch the story unfold with TextWheel: Visualization of large-scale news streams. *Transactions on Intelligent Systems and Technology (TIST)*. ACM. 3(2): 20.

[13] *Cui W., Liu S., Wu Z., Wei H.* 2014. How Hierarchical Topics Evolve in Large Text Corpora. *IEEE Transactions on Visualization and Computer Graphics*. 20(12): 2281–2290.

[14] *Dou W., Yu L., Wang X., Ma Z., Ribarsky W.* 2013. HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies. *Visualization and Computer Graphics, IEEE Transactions*. 19(12): 2002–2011.

[15] *Daud A., Li J., Zhou L., Muhammad F.* 2010. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of computer science in China*. 4(2): 280–301.

[16] *Eades P.* 2010. A heuristic for graph drawing. *Congressus numerantium*. 42: 146–160.

[17] *Gad S., Javed W., Ghani S., Elmqvist N., Ewing T., Hampton N., Ramakrishnan N.* 2015. ThemeDelta: Dynamic Segmentations over Temporal Topic Models. *Visualization and Computer Graphics, IEEE Transactions*. 21(5): 672–685.

[18] *Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P.* 2012. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 3(2): 23.

[19] *Günnemann N., Jarke M.* 2013. D-VITA: A Visual Interactive Text Analysis System Using Dynamic Topic Mining. *BTW workshops*. 237–246.

[20] *Havre S., Hetzler B., Nowell L.* 2002. ThemeRiver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions*. 17(12): 9–20.

[21] *Hofmann T.* 1999. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 50–57.

[22] *Kuhn W.* 1995. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*. 2: 83–97.

[23] *Lau J. H., Grieser K., Newman D., Baldwin T.* 2011. Automatic labelling of topic models. *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA, USA. ACL. 1536–1545.

[24] *Lee H., Kihm J., Choo J., Stasko J., Park H.* 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum*. 31(3): 1155–1164.

[25] *Liu S., Wang X., Chen J., Zhu J., Guo B.* 2014. TopicPanorama: a Full Picture of Relevant Topics. *Visual Analytics Science and Technology (VAST)*. 183–192.

[26] *McCallum A. K.* MALLET: A Machine Learning for Language Toolkit, 2002. `http://mallet.cs.umass.edu`.

[27] *Mei Q., Shen X., Zhai C.* 2007. Automatic labeling of multinomial topic models. *13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA. 490–499.

[28] *Sievert C., Shirley K.* 2014. LDAvis: A method for visualizing and interpreting topics. *Workshop on Interactive Language Learning, Visualization, and Interfaces*.

[29] *Smith A., Hawes T., Myers M.* 2014. Hiérarchie: Interactive Visualization for Hierarchical Topic Models. *Sponsor: Idibon*. 71.

[30] *Snyder J., Knowles R., Dredze M., Gormley M., Wolfe T.* 2013. Topic Models and Metadata for Visualizing Text Corpora. *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 10:5.

[31] *Wei F., Liu S., Song Y., Pan S., Zhou M., Qian W., Shi L., Tan L., Zhang Q.* 2010. TIARA: A Visual Exploratory Text Analytic System. *16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 153–162.