УДК 004.021 + 004.42

ЭВОЛЮЦИЯ МЕТОДОВ ВИЗУАЛИЗАЦИИ КОЛЛЕКЦИЙ НАУЧНЫХ ПУБЛИКАЦИЙ

З. В. Апанович

Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирский государственный *университет, г. Новосибирск*

apanovich@iis.nsk.su

Аннотация

Методы визуализации информации давно зарекомендовали себя как инструмент, позволяющий понимать данные большого объема. Визуализация коллекций научных публикаций является частным случаем визуализации информации. В статье рассмотрены задачи, решаемые при помощи визуализации, модели и методы анализа текстовой информации, а также новые подходы к визуализации документов. Особое внимание уделено тому, каким образом методы визуализации связаны с методами анализа коллекций научных публикаций.

Ключевые слова*: визуализация коллекций документов, анализ текстов, алгоритмы визуализации текстов и метаданных, LDA, NMF, word2vec*

ВВЕДЕНИЕ

В последние годы количество научных публикаций растет экспоненциально. Как известно, в 2006 году было опубликовано 1.35 миллиона научных статей, среднегодовой рост количества публикаций составил 2,5%, и в настоящее время публикуется более 4400 документов в день. Все больше электронных научных коллекций предоставляют полные тексты публикаций. Например, сайт SpringerLink[1] дает исследователям доступ к миллионам научных документов, таким, как книги, статьи в научных журналах и трудах конференций. Все более значительную часть этой коллекции составляют полные тексты публикаций. Доступ к текстам более миллиона публикаций предоставляют такие сервисы, как CEUR Workshop Proceedings[2], библиотека Корнельского университета[3]. Тексты

[1] https://link.springer.com/

[2] http://ceur-ws.org

русскоязычных журнальных публикаций можно найти, например, в научном информационном пространстве Соционет[4]. Коллекции документов являются богатым источником информации. Люди исследуют их, чтобы найти нужные документы, понять содержание коллекций, обнаружить скрытые шаблоны.

Визуализация документов – это класс методов визуализации информации, преобразующих текстовую информацию, такую, как слова, предложения, документы и их взаимоотношения в визуальную форму, сокращая трудозатраты пользователей и позволяя им лучше понимать текстовые документы. По сравнению с другими методами визуализации информации визуализация документов уделяет больше внимания визуализации текстовой информации. В то же время, помимо визуализации информации на текстовом уровне, визуализация документов рассматривает также атрибуты и метаданные документов. Если ранние методы визуализации документов концентрировались больше на визуализации метаданных, таких, как сети цитирования, сети ко-цитирования и сети соавторства [1–6], то с развитием методов анализа и визуализации текстов появляется все больше публикаций, совместно анализирующих и визуализирующих текстовые данные и метаданные. При визуализации коллекций документов принято выделять следующие подзадачи:

— визуализация основного контента коллекции документов,

— визуализация тем коллекции документов,

— визуализация отношений между документами, в частности визуализация сходства документов и кластеризация документов на основе различных оценок сходства,

— визуализация эволюции коллекции документов во времени.

Как правило, эти задачи взаимосвязаны и тесно соседствуют в одной и той же программе визуализации.

Данная статья представляет обзор последних работ, посвященных визуализации коллекций научных публикаций. Особое внимание уделено тому, каким образом методы визуализации связаны с методами анализа коллекций научных публикаций.

[3] https://arxiv.org/

[4] https://socionet.ru/

1. МЕТОДЫ ОБЗОРНОЙ ВИЗУАЛИЗАЦИИ КОЛЛЕКЦИИ ДОКУМЕНТОВ

В последние годы приобрел популярность подход к визуализации коллекций документов, направленный не столько на поиск нужных документов, сколько на представление визуального обзора коллекции документов. Одной из первых систем, реализовавших такой подход, была программа Document Cards [7], которая визуализировала ключевой контент документов как смесь изображений и важных терминов, благодаря чему изображение напоминало карты в карточной игре. На Рис. 1 показана визуализация коллекции документов в виде набора карт, похожих на карточные. На каждой карточке показаны авторы, название статьи, наиболее часто встречающиеся слова в виде облака слов, рисунки, встречающиеся в тексте статьи, а также закладки, позволяющие просматривать текст статьи страница за страницей.

Основным достижением этой работы была демонстрация важности использования изображений в качестве полноправных метаданных, что привело к большому количеству работ, расширяющих этот подход. Так, визуальный обзор, посвященный алгоритмам визуализации деревьев, представлен в работе [8], обзор, рассматривающий визуализацию данных, изменяющихся во времени, представлен в [9], а различные методы визуализации текстовых данных описаны в работе [10].

Отдельно следует остановиться на визуальном обзоре публикаций, касающихся визуализации текстовой информации, представленном в работе [10]. Этот визуальный обзор реализован в виде постоянно пополняющегося сайта [5] и содержит на момент написания данной статьи более 400 названий публикаций, посвященных визуализации текстов. Имеется таксономия, на основе которой можно фильтровать публикации. Эта таксономия содержит такие категории, как аналитические задачи, задачи визуализации, предметные области, которым соответствуют тексты, типы источников данных (документ, корпус или поток текстов), данные и свойства визуальных представлений.

[5] http://textvis.lnu.se/

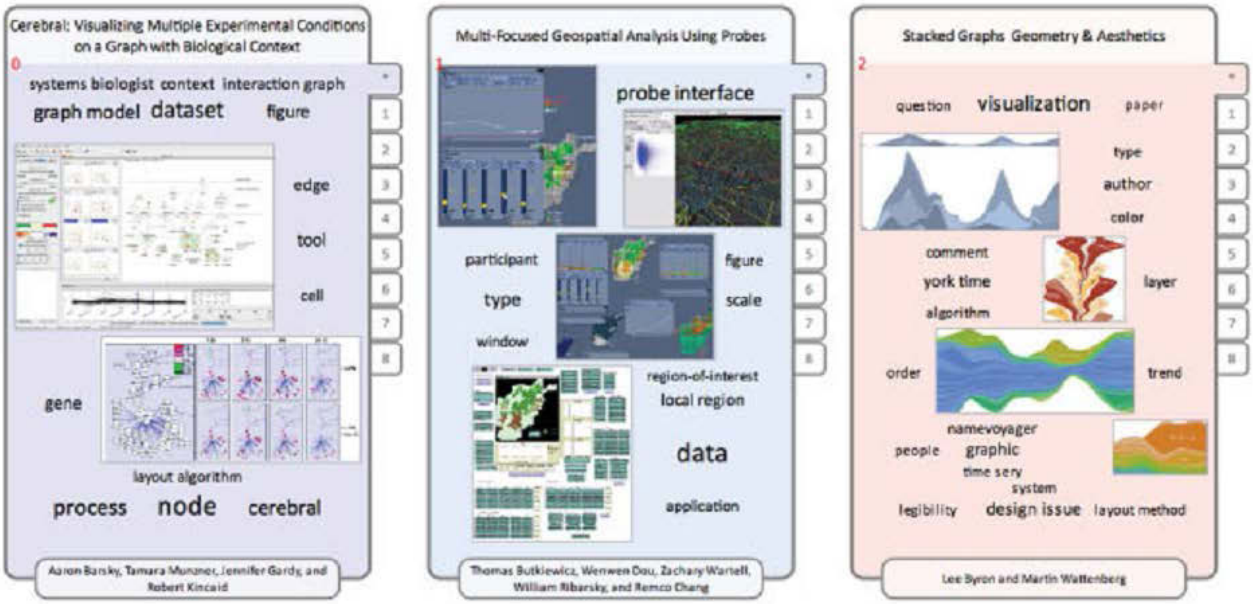


Рис. 1. Визуализация коллекции документов в системе Document Cards [7] в виде набора карт

В силу ограниченности возможностей техники визуального обзора данный вид обзора не может освободить заинтересованного пользователя от необходимости чтения статей, названия которых приведены на сайте, и не подменяет классических обзоров. В частности, на основании визуального обзора и его таксономии невозможно понять, какие именно методы анализа текстовой информации используются в каждой из работ и каким образом методы визуализации связаны с используемыми методами анализа текстов. Этот недостаток мы собираемся устранить в данной работе.

Наконец, в работе [11] представлена система SurVis, позволяющая автоматизировать создание визуальных обзоров коллекций публикаций. Пример визуализации, автоматически создаваемой при помощи этой программы, показан на Рис. 2. Так же, как и в DocumentCards, для каждой публикации создается отдельная «карточка», приводятся основные метаданные публикации, такие, как авторы, название, аннотация, DOI, репрезентативный рисунок из текста публикации (справа). Дополнительно в системе реализовано несколько алгоритмов анализа, позволяющих предоставлять пользователю глобальную информацию обо всей коллекции документов. Имеется временная линия, позволяющая изображать количество публикаций за каждый отдельный период времени, а прямоуголь-

ники, расположенные под временной линией, показывают частоту цитирований (чем темнее прямоугольник, тем больше частота цитирований). При помощи облаков слов резюмируются не только ключевые слова, но и другие метаданные, причем каждый тип метаданных отображается в отдельном облаке слов, где частота каждого термина изображается не только при помощи разных размеров фонтов, но и при помощи численных значений, изображаемых как нижние индексы при каждом термине. Наконец, в системе SurVis реализована возможность кластеризации коллекции документов, как на основе сходства ключевых слов, так и на основе сходства метаданных, таких, как авторы публикаций.

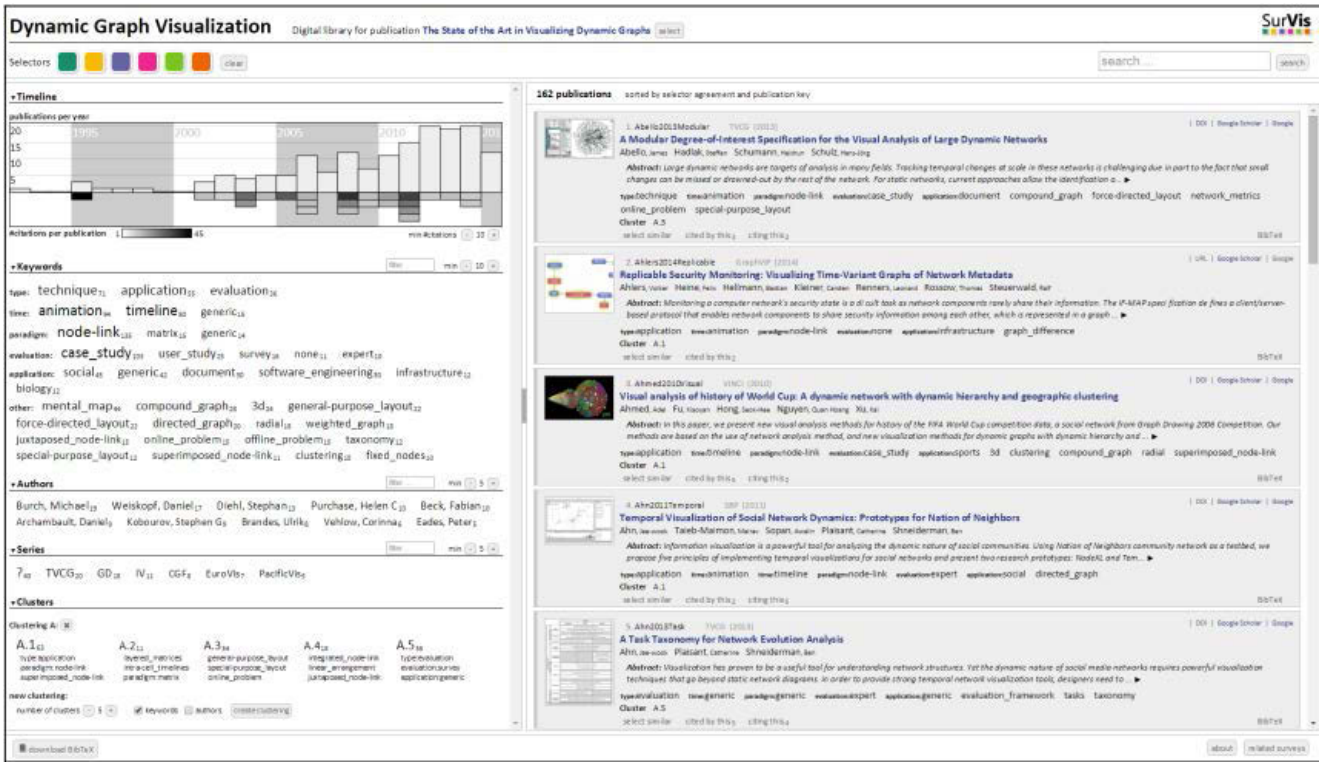


Рис. 2. Survis[11] – программа для создания визуальных обзоров коллекций научных публикаций

Основным достоинством всех визуальных обзоров является достаточно большой список литературы, насчитывающий несколько сотен наименований, что позволяет заинтересованному читателю поработать с рекомендуемым списком литературы и составить собственное представление о предметной области. К недостаткам данного подхода можно отнести то, что не всегда по указанной ссылке доступен полный текст публикации, а рисунок, сопровождающий описание каждой статьи, как правило, достаточно сложно интерпретировать, не про-

читав текст исходной статьи. Наконец, данные методы не предназначены для работы с коллекциями большого объема, создаваемыми автоматически, а не вручную. Методы визуализации больших коллекций документов будут описаны в последующих разделах.

2. МОДЕЛИ ПРЕДСТАВЛЕНИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ, ИСПОЛЬЗУЕМЫЕ ПРИ ВИЗУАЛИЗАЦИИ КОЛЛЕКЦИЙ ДОКУМЕНТОВ

В основе методов визуализации документов лежат различные модели представления текстов. Пусть D – множество (коллекция) документов, W – множество (словарь) всех употребляемых терминов. Каждый документ $d \in D$ представляет собой последовательность терминов (w_1, w_2 ,..., w_N) из словаря W. Термины могут повторяться в документе несколько раз. В простейшем случае векторная модель сопоставляет каждому документу точку в многомерном пространстве, где каждое измерение соответствует одному термину, а вклад каждого термина пропорционален его весу в документе.

Вес термина в документе часто вычисляется при помощи меры TF-IDF, пропорциональной количеству употреблений термина в документе и обратно пропорциональной частоте употребления термина в других документах коллекции:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) = f_{t,d} \times \log \frac{|D|}{n_t + 1},$$

где $|D|$ – общее количество документов, n_t – количество документов, содержащих терм t. В качестве меры близости между двумя документами, заданными векторами $X = (x_1, x_2, \ldots, x_N)$ и $Y = (y_1, y_2, \ldots, y_N)$, чаще всего используют косинусную меру близости:

$$\cos(x, y) = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}}.$$

Примером системы, использующей представление TF-IDF для визуализации коллекции документов, является рассмотренная выше система SurVis [11]. Недостатком этой модели является высокая размерность векторного пространства. Даже после применения стандартных методов предварительной обработки, таких, как удаление стоп-слов, удаление наиболее часто встречающихся слов

и наиболее редко встречающихся слов, а также лемматизации, приводящей каждое слово к его нормальной форме, или стемминга, отбрасывающего изменяемые части слов, размерность векторного пространства остается большой. Поэтому применяются другие модели, уменьшающие размерность векторного пространства.

Одна из моделей, которая существенно снижает размерность векторного представления текста, носит название «*вероятностное моделирование тем».* Вероятностное моделирование тем – это множество статистических методов обработки корпуса документов, которые идентифицируют скрытые темы в корпусе документов. Каждый документ моделируется как вектор скрытых тем с весами, а каждая тема моделируется как вектор слов с весами, встречающихся в одном и том же документе. *Вероятностная скрытая семантическая индексация* (probabilistic latent semantic indexing, pLSI) [12] и *скрытое размещение Дирихле* (Latent Dirichlet Allocation, LDA) [13] – два популярных метода в этой категории. Метод pLSI основан на принципе максимума правдоподобия, а метод LDA предполагает, что распределение тем в документах и распределение слов по темам априори имеют распределения Дирихле. Примером использования представления LDA для визуализации коллекций документов являются работы [14–18]. В случае больших коллекций документов с большим количеством тем их принято организовывать иерархически с применением модели «*Байесовское розодерево»* (Bayesian Rose Tree, BRT) [19], которая использует байесовский алгоритм иерархической кластеризации для построения дерева тем с произвольным коэффициентом ветвления, в котором каждая не листовая вершина изображает кластер тем. Примерами систем визуализации, использующих иерархическое представление моделей тем, являются TopicPanorama [20] и RoseRiver [21]. Однако общим недостатком всех вариаций модели LDA является высокие требования к производительности. С недавнего времени, *неотрицательное матричное разложение* (non-negative matrix factorization, NMF), в котором матрица термин-документ представляется в виде произведения двух матриц с неотрицательными элементами [22], используется в качестве альтернативного подхода к моделированию тем в анализе документов. Поскольку модель NMF имеет существенно более высокую скорость работы по сравнению с LDA, на ее основе разработано несколько систем, динамически управляющих процессом моделиро-

вания тем. Примером использования представления NMF для визуализации коллекций документов являются программы Utopian [23] и TopicLens [24]. Так, в программе Utopian [23] возможные взаимодействия с коллекцией документов включают уточнение темы путем изменения веса ключевых слов в теме, слияние похожих тем, разделение тем и создание тем на основе выбранных пользователем документов или ключевых слов.

Наконец, еще одна модель представления текста, порождающая векторные пространства относительно небольшой размерности, стала чрезвычайно популярной в последние годы. Программный инструмент word2vec [25] основан на дистрибутивной семантике и векторном представлении слов. *Векторное представление слов* основано на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (и, значит, имеющие схожий смысл), в векторном представлении будут иметь близкие координаты. В word2vec существуют два основных алгоритма обучения: CBOW (непрерывный мешок слов) и Skip-gram. Архитектура CBOW предсказывает текущее слово, исходя из окружающего его контекста. Архитектура Skip-gram использует текущее слово, чтобы предугадывать окружающие его слова [25]. Авторы этой модели утверждают, что она дает лучшие результаты, чем pLSA, а по скорости работы превосходит LDA. Примером использования представления word2vec для визуализации коллекций документов является работа cite2vec [26].

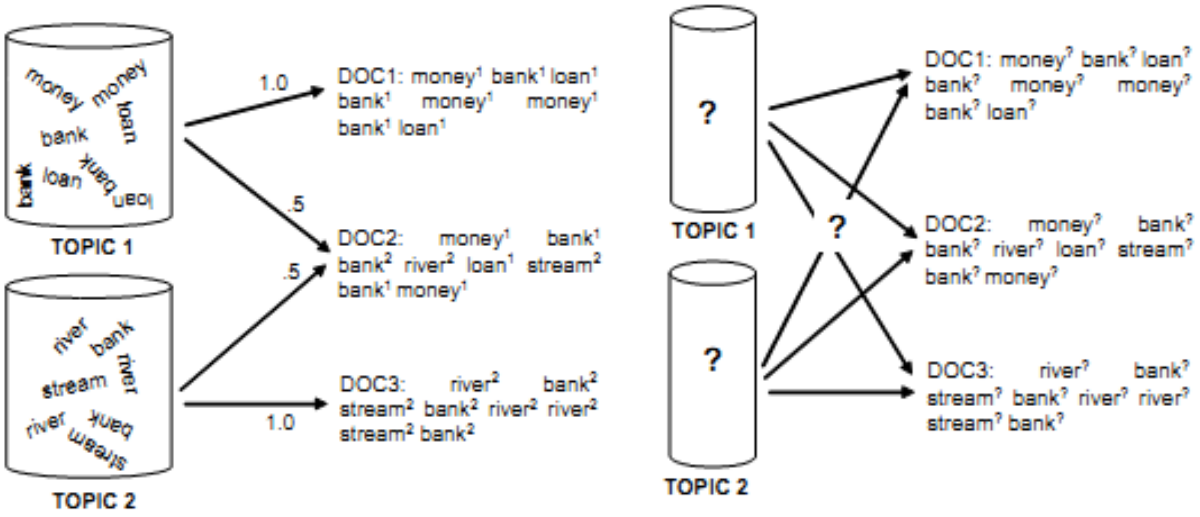
3. ВЕРОЯТНОСТНОЕ МОДЕЛИРОВАНИЕ ТЕМ И ПРИМЕНЕНИЕ ЭТОЙ МОДЕЛИ ДЛЯ ВИЗУАЛИЗАЦИИ КОЛЛЕКЦИЙ ДОКУМЕНТОВ

В основе методов вероятностного моделирования тем лежит предположение о том, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна. Коллекция документов рассматривается как множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$. Документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, тема $t \in T$ является *латентной* (скрытой) переменной.

Порождающая модель вероятностного моделирования определяет простую вероятностную процедуру, которая описывает, как слова в документах могут создаваться на основе скрытых (случайных) переменных. Чтобы создать новый документ, надо выбрать распределение по темам. Тогда для добавления каждого нового слова в этот документ случайным образом выбирается тема в соответствии с этим распределением, и затем извлекается слово из этой темы.

Чтобы инвертировать этот процесс, то есть построить тематическую модель коллекции документов, надо найти множество тем T, распределения $p(w|t)$ для всех тем $t \in T$ и распределения $p(t|d)$ для всех документов $d \in D$. Это значит, что надо найти наилучший набор скрытых переменных, которые могут объяснить наблюдаемые данные (т. е. наблюдаемые слова в документах) в предположении, что эти данные были сгенерированы при помощи этой порождающей модели. На Рис. 3 показан пример [27] порождающего процесса и процесса вероятностного моделирования. Предположим, что коллекция документов порождается на основе двух тем (TOPIC1 и TOPIC2). Тема 1 связана с понятием «деньги», а тема 2 – с понятием «река». Поэтому одно и то же слово bank имеет в этих темах разный смысл. Если в теме 1 слово bank понимается как банк, в котором хранятся деньги, то во второй теме оно имеет смысл «берег реки».

Процедура порождения коллекции документов состоит в выборе различных слов из темы в зависимости от веса, заданного каждой теме. Например, документ DOC1 был получен путем отбора слов только из темы 1, а документ DOC3 – выбором слов только из темы 2, в то время как документ DOC2 был сформирован сочетанием двух тем в равных пропорциях. Надстрочные номера, связанные со словами в документах, указывают, какая из тем была использована для выборки слова.



(а) (б)

Рис. 3. (а) Пример порождения коллекции документов при помощи слов, выбираемых из двух тем. (б) Вычисление распределений тем в документах при помощи вероятностного вывода [27]

В данном определении модели нет требования взаимной исключительности тем, которая позволяла бы словам быть частью только одной темы. Это позволяет тематическим моделям описывать многозначность, когда одно и то же слово может иметь несколько значений. Например, обе темы, и первая тема, связанная с деньгами, и вторая, связанная с реками, могут с высокой вероятностью содержать слово “bank”, что разумно, учитывая многозначный характер этого слова. На Рис. 3(б) показан пример задачи вероятностного вывода, решаемой для того, чтобы определить распределения тем в коллекции документов.

На Рис. 4 показан пример четырех тем, выделенных из коллекции документов при помощи вероятностного вывода [27]. Как правило, темы, вычисленные при помощи метода LDA, обладают свойством хорошей интерпретируемости, то есть слова, принадлежащие одной теме, легко обобщить. Так, например, тема 247 объединяет слова, касающиеся лекарств, тема 5 – цветов, тема 43 – мышления, а тема 56 – больниц. Темы можно сравнивать между собой как вектора, например, при помощи мер косинусной близости или дивергенции Кульбака –Лейблера.

За последние пятнадцать лет реализовано большое количество экспериментальных систем, использующих вероятностное моделирование тем для анализа и визуализации коллекций документов большого объема. Лучше всего данные методы подходят для реферирования коллекций документов, определения сходства между документами и кластеризации на основе сходства, а также для изображения эволюции во времени коллекций документов.

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Рис. 4. Пример четырех тем, выделенных из коллекции документов при помощи вероятностного вывода [27]

Наиболее очевидный способ визуализации результатов тематического моделирования состоит в непосредственном изображении матрицы тем и матрицы документов так, что численные значения вероятности темы в данном документе или вероятности слова в данной теме изображаются при помощи глифов, например, окружностей, радиус которых пропорционален значению этих вероятностей. В работе Serendip [14] визуализация использована для изображения распределения тем в документах, а в работе Termite [15] – для изображения распределения слов по темам.

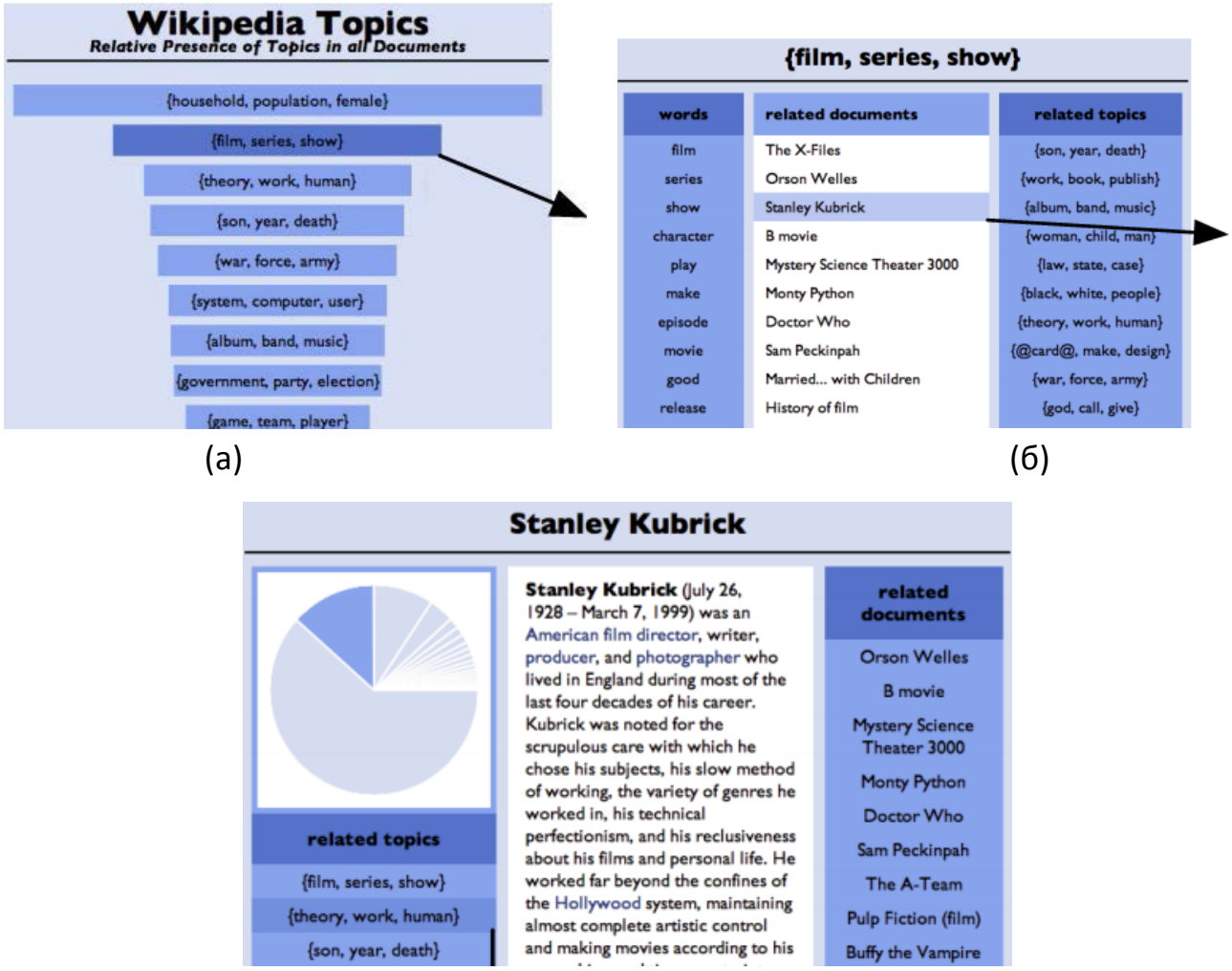
Более продвинутый подход к визуализации на основе LDA представлен в работе [16]. В ней представлен Навигатор, позволяющий перемещаться по большой коллекции документов на основе их тематического содержания. Для

просмотра коллекции в Навигаторе используется три типа страниц, примеры которых показаны на Рис. 5:

— стартовая страница Навигатора (Рис. 5(а)),

— страница темы (Рис. 5(б)),

— страница документа (Рис. 5 (с)).



(а) (б)

(с)

Рис. 5. Основные типы страниц Навигатора на основе метода LDA. (а) Стартовая страница Навигатора. (б) Страница темы. (с) Страница документа [16]

Навигация в коллекции документов (в данном случае, коллекции статей Википедии) начинается со стартовой страницы, где показан список тем, извлеченных из коллекции документов при помощи метода LDA. Каждая тема представлена прямоугольником, содержащим три наиболее значимых слова данной

темы. Ширина прямоугольника изображает распределение тем в коллекции документов. При выборе одной из тем Навигатор переходит на страницу выбранной темы. (Рис. 5(б)). Страница темы разделена на три столбца. Слева – список слов, упорядоченный по их распределению вероятностей в данной теме, в центре – документы, связанные с этой темой и упорядоченные по значимости данной темы в документе. Справа – связанные темы, в качестве которых выдаются темы, имеющие аналогичное распределение. При выборе одного из документов на странице темы Навигатор переходит на страницу выбранного документа (Рис. 5(с)). Страница документа показывает документ и темы, связанные с этим документом и упорядоченные по важности тем. В левом столбце вверху изображена круговая диаграмма тем, присутствующих в документе. В правом столбце – другие сходные документы. Сходство между двумя документами определяется по специальной формуле, говорящей, что похожие документы имеют сходные комбинации тем.

Хотя такая визуализация является достаточно понятной, она не отражает глобальных отношений между темами и документами. Для визуализации глобальных отношений между темами и документами используются другие алгоритмы.

4. ВИЗУАЛИЗАЦИЯ ГЛОБАЛЬНЫХ ОТНОШЕНИЙ МЕЖДУ ТЕМАМИ И ДОКУМЕНТАМИ

Как уже упоминалось ранее, результаты тематического анализа можно рассматривать как «жесткую», так и «мягкую» кластеризацию слов и документов по темам. В случае жесткой кластеризации каждый документ *i* назначается ровно одной теме *j*, так что

$$j = \underset{j \epsilon T}{\operatorname{argmax}} \theta_{ij} ,$$

где *T* – множество всех тем, а θ_{ij} – вероятность наличия темы *j* в документе *i.* Таким образом, все множество документов разбивается на кластеры, где количество кластеров равно количеству тем. На Рис. 6 показаны результаты «жесткой» кластеризации коллекции документов на основе LDA, реализованной в программе iVisClustering [17]. Для размещения узлов графа, соответствующих отдельным документам, используется обычный силовой алгоритм. В этом представлении кластеры-темы хорошо различимы визуально, позволяя легко понять

общую структуру данных. Каждый узел-документ изображается как цветной кружок. Узлы документов одного и того же цвета относятся к одному и тому же кластеру-теме. Ребра между узлами представляют собой сходство документов на основе косинусной близости. Каждому кластеру также соответствует цветная прямоугольная рамка, в которой размещено пять слов, имеющих наивысшую вероятность для данной темы. Имеются еще невидимые ребра, которые связывают узел-кластер с узлами-документами, входящими в этот кластер. Изменяя длину пружины ребра, можно управлять размером каждого кластера. Если длина пружины ребра установлена на небольшое значение, все узлы, принадлежащие одному кластеру, собираются в одну точку.

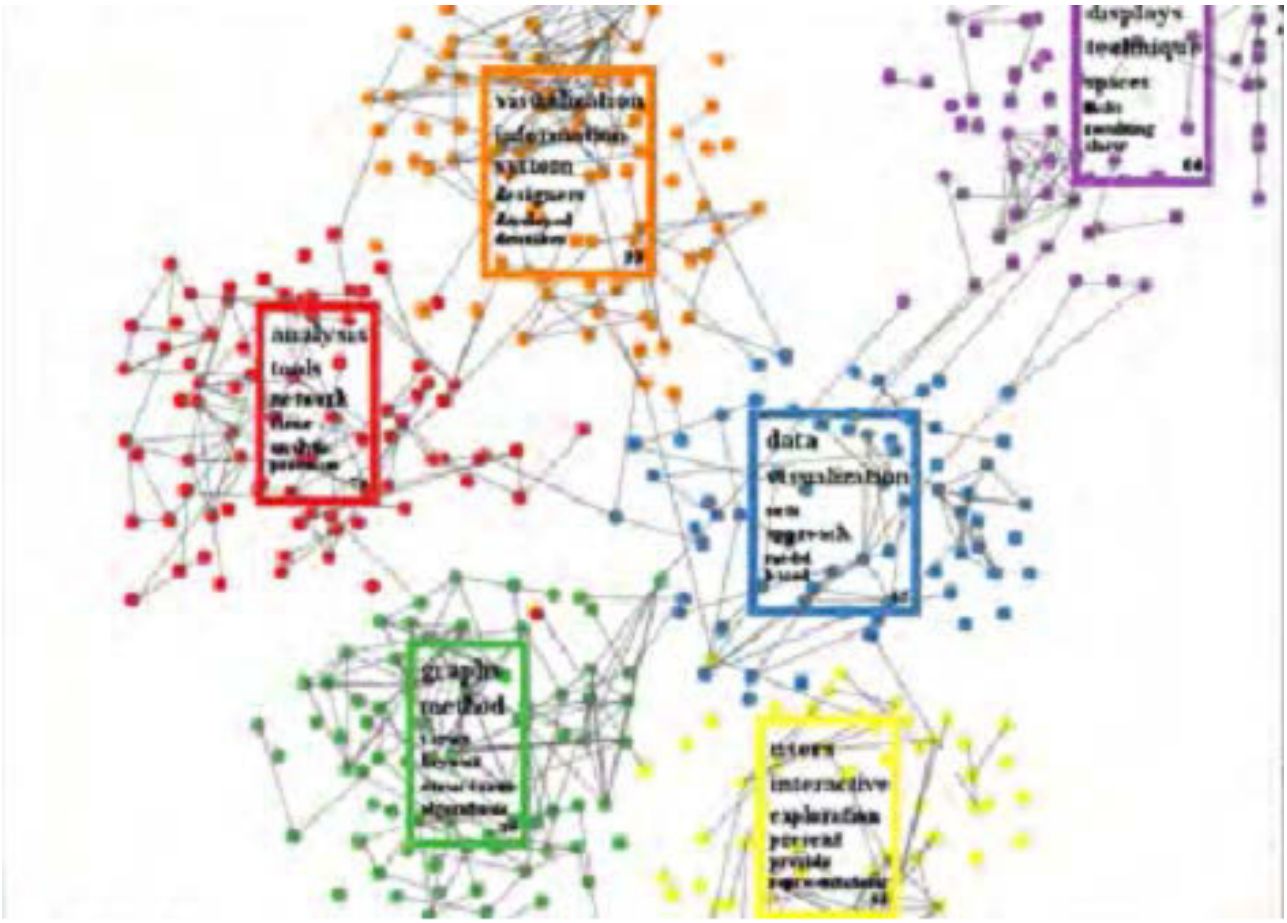
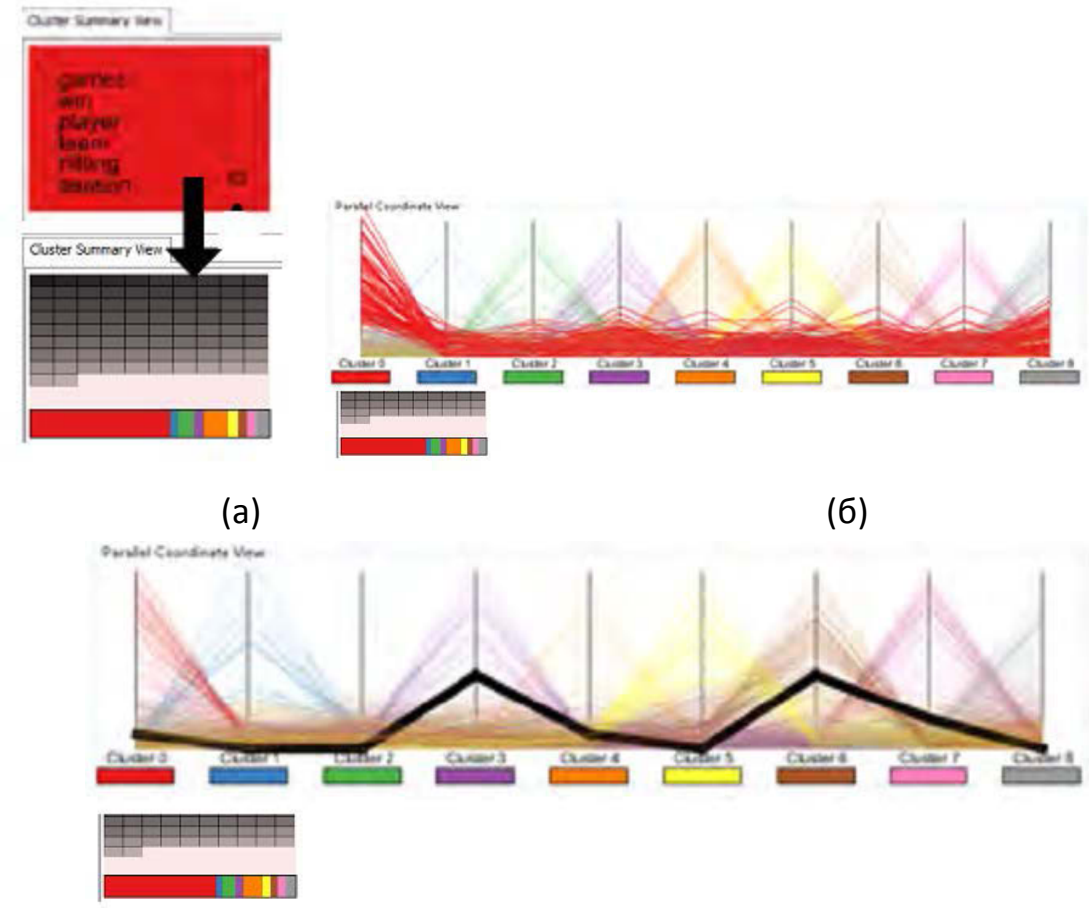


Рис. 6. Результаты «жесткой» кластеризации коллекции документов на основе LDA [17]

При наведении мыши на один из кластеров в программе включается так называемый «рентгеновский режим», при котором можно просмотреть не только результаты «жесткой» кластеризации, но и результаты «мягкой» кластериза-

ции. Рентгеновский режим отображает документы на сетке, и каждый квадрат сетки соответствует одному документу, входящему в текущий кластер (Рис. 7(а)). Квадрат сетки темный, если соответствующий документ сильно связан с данной темой, и светлый, если он с ней связан слабо. Рентгеновский режим также содержит цветовой спектр, расположенный ниже квадратов сетки документа, который показывает, как данный кластер связан с другими кластерами. В этом цветовом спектре каждый цвет соответствует одному кластеру, поэтому связь между выбранным кластером и другими кластерами можно идентифицировать, наблюдая ширину цветной области.



(а) (б)

(в)

Рис. 7. Визуализация результатов «мягкой» кластеризации коллекции документов. (а) Изображение документов, принадлежащих одному кластеру. (б) Изображение кластеров в параллельных координатах. (в) Выделение одного документа на изображении в параллельных координатах [17]

При перемещении мыши на определенный квадрат сетки, соответствующий документ выделяется в Параллельных координатах толстой черной линией, как показано на рисунке 7(в). Цветовая шкала в нижней части сетки документов показывает взаимосвязь между выбранным документом и кластерами. Это взаимодействие между «рентгеновским» представлением и представлением в Параллельных Координатах позволяет исследовать данные различными способами. Например, перемещая курсор мыши над сеткой в рентгеновском представлении, можно наблюдать шаблоны отдельных документов в режиме параллельных координат. Это позволяет быстро найти документы, которые содержат несколько тем.

В представлении Параллельные координаты для каждого документа *i* визуализируется вектор $\vartheta_i = (\vartheta_{i1}, \vartheta_{i2}, ..., \vartheta_{ik})$, который представляет собой результат мягкой кластеризации коллекции документов по *k* темам, так же, как это сделано в работе ParallelTopics [18]. Каждая вертикальная ось соответствует одной теме, и точка на этой оси соответствует значению вероятности присутствия соответствующей темы в документе *i*. Каждая ломаная линия параллельных координат соответствует одному документу и кодируется цветом в зависимости от членства документа в кластере.

Представление коллекции в параллельных координатах может взаимодействовать с другими изображениями, в частности, с жестким представлением кластеров. Когда курсор мыши помещается на узел кластера, активизируется «рентгеновский» режим, а документы соответствующего кластера выделяются в виде параллельных координат. Цель этого взаимодействия – дать пользователям понять или общую структуру распределения тема-документ в кластере или характеристики одного документа. Например, ломаная линия с несколькими пиками, изображающая документ в параллельных координатах, указывает на то, что документ представляет собой смесь связанных между собой тем. Кроме того, если большинство документов в кластере имеет одинаковые множественные пики, это указывает на то, что темы с пиками связаны друг с другом. Визуализация в режиме параллельных координат также имеет ползунок для установки порогового значения. Если задано пороговое значение, оно удаляет документы со

значениями ϑ_{ij} ниже порогового значения по всем темам, чтобы убрать «зашумляющие» документы, которые явно не относятся к определенному кластеру.

Наконец, от представления на уровне кластеров можно получить доступ к текстам отдельных документов, что позволяет понять, почему тот или иной документ присваивается определенному кластеру-теме. Изображение отдельного документа в системе IvisClustering показано на Рис. 8. В этом представлении наряду с исходным текстом представлена раскраска терминов разными цветами, в соответствии с тем, к каким темам-кластерам относится тот или иной термин.

Рис. 8. Цветовое изображение тем в отдельно взятом документе [17]

Представление документа состоит из трех частей: фактического текста документа (в центре), списка документов (слева) и ползунка для управления значением *k*, равным количеству высвечиваемых ключевых слов одной темы. Если установить *k* на небольшое значение, можно увидеть наиболее важные слова, а затем, постепенно увеличивая значение *k,* можно увидеть, какой цвет доминирует в документе. Если определенный цвет доминирует в этом изображении, это означает, что выбранный документ сильно связан с соответствующей темой.

5. ВИЗУАЛИЗАЦИЯ ЭВОЛЮЦИИ КОЛЛЕКЦИИ ДОКУМЕНТОВ ВО ВРЕМЕНИ НА ОСНОВЕ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Если документы в коллекции имеют временные штампы, то появляется возможность визуализации и визуального исследования эволюции тем в коллекции документов. Графы, изображающие временные ряды при помощи смежных слоев, появились достаточно давно. Одной из первых работ, относящихся к этому классу визуализаций, можно считать работу ThemeRiver [28], которая строила гладкую интерполяцию дискретных данных, при этом ось *x* являлась центральной линией изображения, вдоль которой симметрично располагались слои. Этот алгоритм визуализации был использован в нескольких системах

визуализации коллекций документов, среди них одной из самых известных является Tiara [29].

На Рис. 9 показана визуализация эволюции тем в коллекции документов в системе Tiara. Каждый цветной слой соответствует одной извлеченной теме, реферируя контент темы и эволюцию контента во времени. По оси *x* изображены документы, упорядоченные по времени, ширина слоя (по оси *y*) изображает силу темы, измеряемую как количество документов, соответствующих определенной теме в заданный момент времени. Аналогично слова, расположенные в одном и том же слое, соответствуют наиболее часто встречающимся словам данной темы в разные моменты времени. Данный способ изображения хорошо подходит для изображения статической коллекции документов, с небольшим количеством тем. В то же время в большой коллекции документов количество тем может измеряться десятками и сотнями. Стало быть, для визуализации больших коллекций документов необходимо разрабатывать и исследовать методы иерархической визуализации.

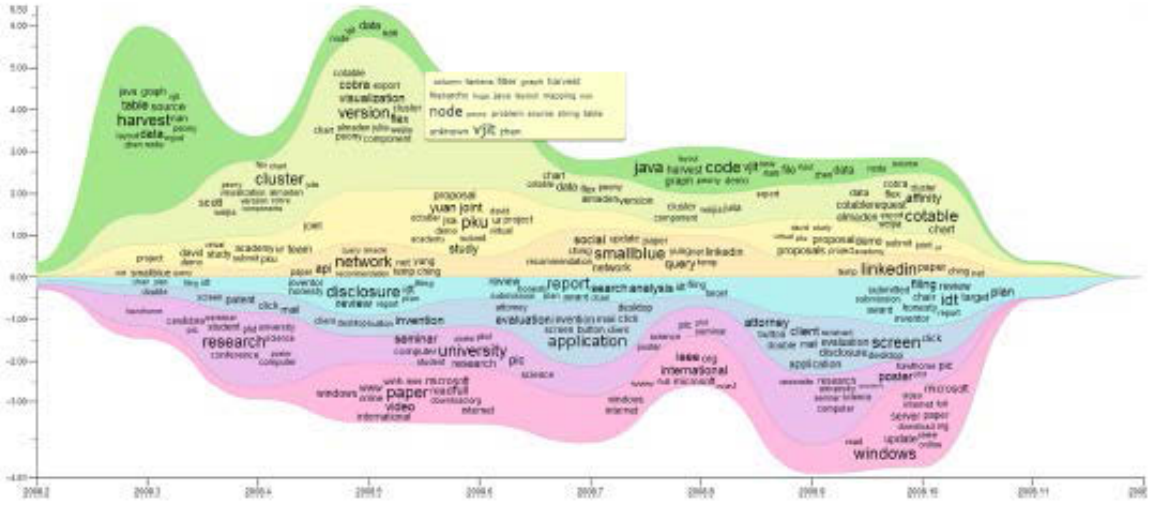


Рис. 9. Визуализация эволюции тем в коллекции документов в системе Tiara [29]

Также в случае, если коллекция допускает изменение во времени, с течением времени могут возникать новые темы, а старые – исчезать. При иерархической визуализации некоторые темы могут со временем объединяться, а некоторые, наоборот, разделяться на подтемы. Различные системы визуализации используют разные модели иерархической кластеризации для выделения дерева тем. Такие программы, как TopicPanorama[20] и RoseRiver [21], используют байе-

совское розо-дерево (Bayesian Rose Tree, BRT) [19] для построения статической иерархии тем из множества тем, но их метод не работает с динамическими коллекциями документов. Метод, реализованный в системе RoseRiver, позволяет сгенерировать множество эволюционирующих деревьев тем, которое используется, чтобы гладко отобразить большое количество тем, изменяющихся во времени.

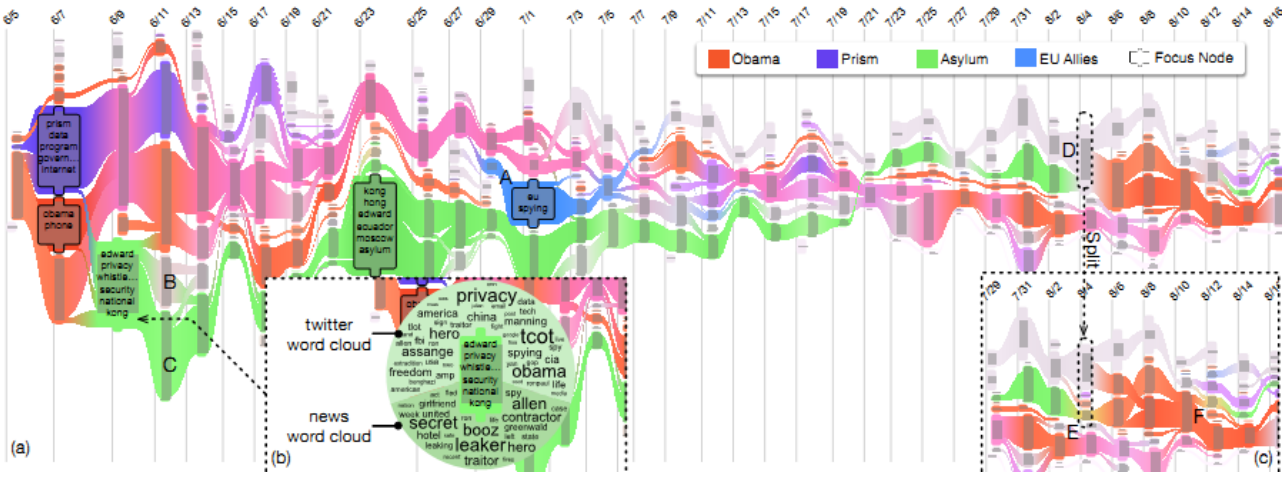


Рис. 10. Визуализация эволюционирующих тем, организованных иерархически, реализованная в системе RoseRiver [21]

На Рис. 10 можно видеть пример такой реализации, показывающей изменение коллекции документов день за днем в течение трех месяцев. Показано, как тема, отмеченная стрелкой (4 августа), разделяется на две подтемы. Специфической особенностью такого способа визуализации является то, что каждому моменту времени соответствует иерархическое представление тем, представляющее собой дерево небольшой глубины (три–четыре уровня). Поскольку изображение поддерева, соответствующего каждому временному интервалу, требует использования нескольких вертикальных столбцов, пространство экрана не может вместить большой временной период, а сам алгоритм работает достаточно медленно, что затрудняет работу пользователя с данной визуализацией. Более того, возникает необходимость в другом алгоритме визуализации, который мог бы показать разделение и слияние тем. В качестве такого алгоритма можно использовать вариант известного метода поуровневого изображения ориентированных графов [30].

Общим недостатком всех вариаций модели LDA являются высокие требования к производительности. Хотя выше перечисленные системы являются ин-

терактивными, они имеют тенденцию статического использования результатов моделирования тем, так как темы вычисляются один раз и не перевычисляются повторно. В то же время, для повышения качества представления данных может возникать потребность в улучшении результатов тематического моделирования в процессе взаимодействия с пользователем. Чтобы обеспечить возможность повторного вычисления тем в реальном времени, необходимо либо использовать более высокопроизводительные алгоритмы, либо более высокоскоростные компьютеры, либо и то и другое. Примером использования параллельных вычислений для LDA является система TopicNets [31]. Она позволяет итеративно пересчитывать результаты моделирования тем LDA на динамически изменяющемся подмножестве документов, просматриваемых пользователем. В то же время появился ряд систем, использующих для анализа данных более эффективные алгоритмы, в частности, различные варианты алгоритма неотрицательной матричной факторизации.

6. ИНТЕРАКТИВНЫЕ ВИЗУАЛИЗАЦИИ КОЛЛЕКЦИЙ ДОКУМЕНТОВ НА ОСНОВЕ АЛГОРИТМА НЕОТРИЦАТЕЛЬНОГО МАТРИЧНОГО РАЗЛОЖЕНИЯ

Стандартный алгоритм неотрицательного матричного разложения (NMF) выполняет моделирование тем следующим образом.

Пусть дан набор документов, представленный в виде матрицы терм-документ $X \in R_+^{mxn}$, который содержит *n* документов, состоящих из *m* слов. Для заданного количества тем, такого, что $k<<\min\{m, n\}$, стандартный NMF вычисляет низкоуровневое приближение матрицы *X* произведением двух неотрицательных матриц *W* и *H*, где $W \in R_+^{mxk}$ и $H \in R_+^{kxn}$:

$$\min_{W,H\geq 0} \|X - WH\|_F^2.$$

Для оценки аппроксимации используется норма Фробениуса, которая для произвольной матрицы *A* вычисляется по формуле

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \ .$$

Одна из результирующих матриц *W* представляет собой набор из *k* тем, и каждый ее столбец соответствует одной теме, представленной как взвешенная комбинация из *m* терминов. Чем больше вес элемента w_{ij}, тем более релевантным считается соответствующее слово *j* теме *i*. Точно так же матрица *H* представляет собой набор из *n* документов, где каждый из столбцов соответствует одному документу и описывается как взвешенная комбинация *k* тем. Чем больше вес элемента h_{ij}, тем более релевантным считается соответствующий документ *i* теме *j*. В терминах «жесткой» кластеризации тема, связанная с наибольшим значением в каждом столбце матрицы *H*, определяет членство соответствующего документа в кластере тем.

Поскольку модель NMF имеет существенно более высокую скорость работы по сравнению с LDA, на ее основе разработано несколько систем, позволяющих динамически управлять процессом моделирования тем. Так, в системе UTOPIAN [23] реализовано несколько нетривиальных возможностей управления моделированием тем, таких, как уточнение темы путем изменения веса ключевых слов в теме, разделение и слияние тем, создание новой темы на основе выбранных пользователем документов или ключевых слов.

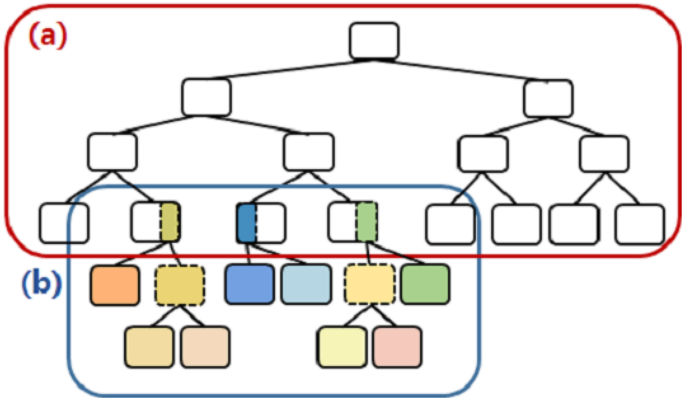
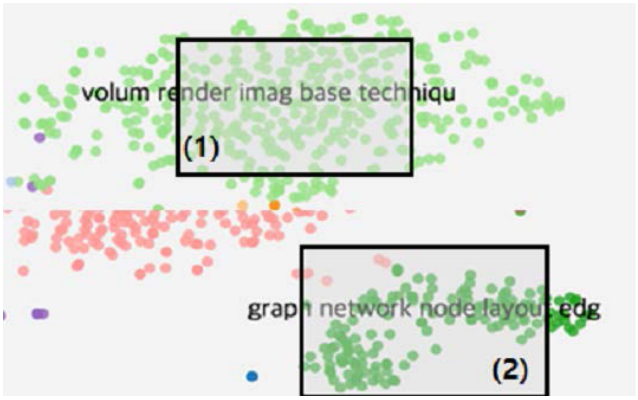


Рис. 11. (а) Исходное бинарное дерево тем, (б) Бинарное дерево тем, динамически сгенерированное для подмножества документов, выделенного при помощи «лупы» [24]

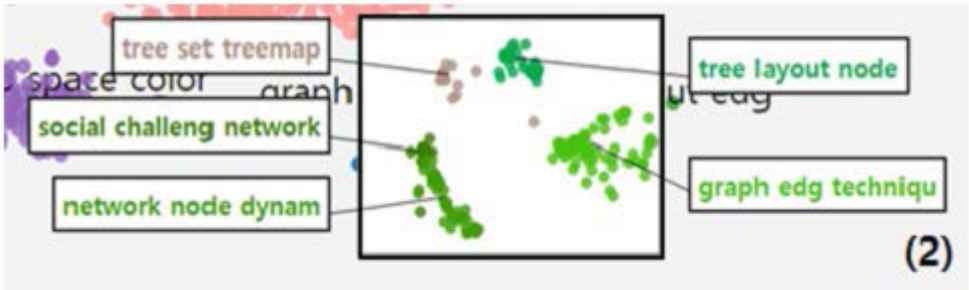
Еще более эффективен иерархический алгоритм неотрицательного матричного разложения (H-NMF), описанный в работе [22]. Этот алгоритм выполняет иерархическую кластеризацию коллекции документов путем построения двоичной иерархии тем, как показано на Рис. 11(а). При этом множество документов так же разбивается на две группы, соответствующие двум новым дочерним

узлам дерева. Если надо получить *k* тем, рекурсивный процесс расщепления H-NMF продолжается до тех пор, пока общее число листовых узлов в двоичном дереве тем не станет равным *k*. Критерий для определения того, какой узел будет разбит на следующем шаге, основан на специальной оценке, которая измеряет, насколько отличаются две новых созданных темы (соответствующие двум дочерним узлам) от темы их родительского узла. Доказано, что H-NMF работает значительно быстрее, чем стандартный NMF, при генерации такого же количества тем.

Развитие иерархического подхода к выделению тем и интерактивной кластеризации коллекции документов представлено в работе TopicLens [24]. В ней реализована метафора «лупы», которая при просмотре больших текстовых коллекций позволяет пользователю в интерактивном режиме выделять группы публикаций, наиболее точно соответствующих его интересам. При перемещении лупы динамически перевычисляются как темы на тех документах, которые попали в зону действия лупы, так и их проекции на плоскость, показывая более мелкозернистую структуру коллекции документов. Как показано на Рис 11 (a), первоначально в коллекции документов выделено два тематических кластера, раскрывающих темы, связанные с такими терминами как “volume,” “render,” “graph,” и “network.” Для дальнейшего изучения подробной информации, связанной с этими исследовательскими областями визуализации, лупа TopicLens была применена к области (2).



(a)



(б)

Рис. 12. Лупы TopicLens [24]. (а) Начальное моделирование тем. (б) Уточнение тем для области 2

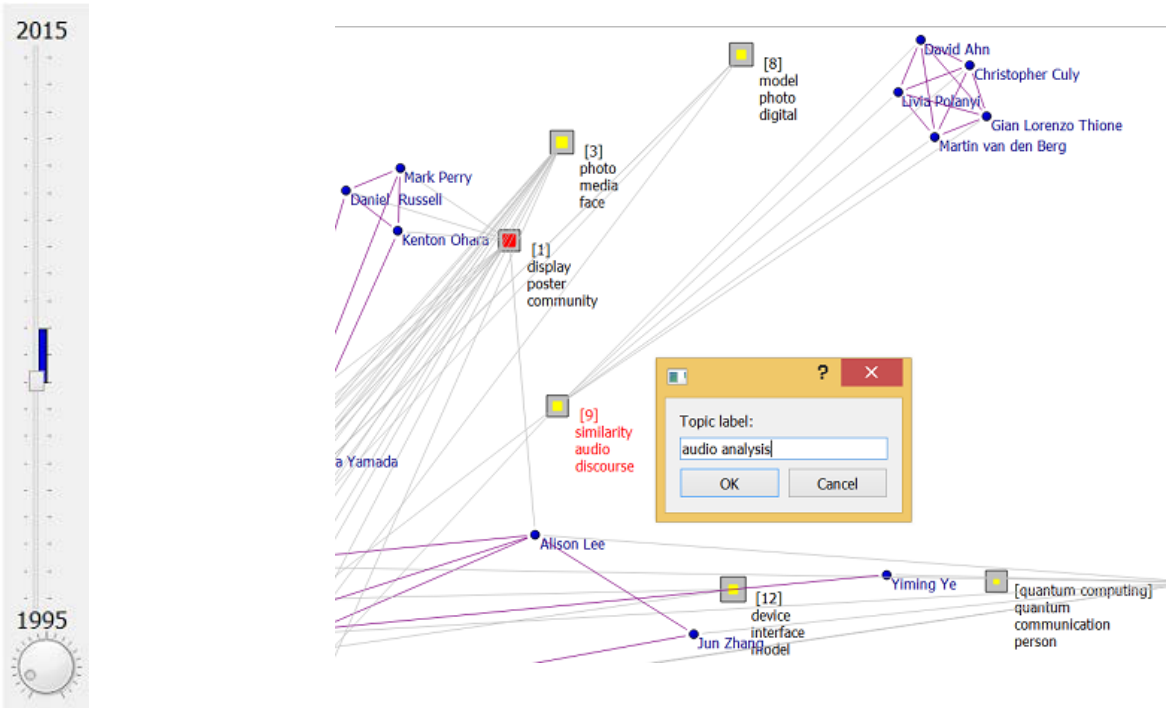
На Рис. 12 (б) показан результат применения лупы TopicLens к области (2). Можно видеть, что после применения лупы TopicLens к этой области на изображении появились новые значимые термины, такие, как “social” и “tree”, которые соответствуют таким областям исследований, как размещение деревьев и социальная сеть. При изучении деталей документов, попавших в соответствующий узел бинарного дерева, было обнаружено несколько статей, соответствующих выделенным терминам, например, в узле дерева была обнаружена статья под названием “Using SocialAction to uncover structure in social networks over Time.” Таким образом, было продемонстрировано, что реализованная иерархическая кластеризация эффективно выделяет темы более низких уровней.

7. СИСТЕМЫ ВИЗУАЛИЗАЦИИ, ИНТЕГРИРУЮЩИЕ ВИЗУАЛИЗАЦИЮ ТЕКСТОВ И ВИЗУАЛИЗАЦИЮ АТРИБУТОВ

В то время как в предыдущих разделах рассматривались возможности визуализации коллекции документов на основе анализа текстового контента, в последнее время появляется все больше систем, использующих при визуализации как результаты анализа текстового контента, так и метаданных публикаций. В данном разделе будут рассмотрены три системы этого класса. В работе [32] анализ текста используется для более детальной визуализации тем, являющихся предметом научного сотрудничества (соавторства) различных исследователей. А в работах CiteRivers [33] и Cite2vec [26] информация о цитировании научных публикаций интегрирована с визуализацией текстового контента.

В отличие от стандартных методов визуализации сотрудничества на основе сетей соавторства, где основными сущностями являются авторы и публикации, в

работе [32] основными визуализируемыми сущностями являются авторы и темы. Метод LDA используется для автоматического вычисления тем совместных публикаций на основе метаданных научных публикаций, таких, как название, авторы, ключевые слова и абстракт. На Рис. 13 показан пример визуализации, создаваемой этой программой. Темы, извлеченные из документов, изображаются в виде фиксированных вершин квадратной формы, а авторы публикаций, работающие над разными темами в разные моменты времени, изображаются перемещаемыми вершинами круглой формы. Важным элементом визуализации является ползунок, позволяющий выбирать произвольный временной интервал и исследовать отношения сотрудничества в выбранном интервале времени. Каждая вершина-тема помечена тремя наиболее важными для нее словами, и цвет каждой вершины соответствует публикационной активности по этой теме в заданном временном интервале. Выбор временного интервала используется для фильтрации узлов-авторов: только авторы, имеющие публикации в выбранном временном интервале, появляются на визуализации.

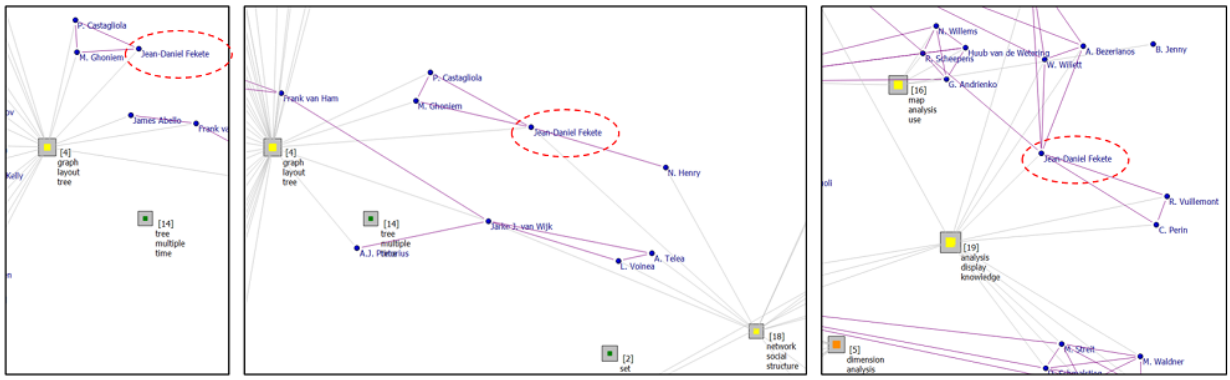


(а) (б)

Рис. 13. Связи между авторами, темы над которыми они сотрудничают, уровни всплесков тем за выбранный период времени [32]

Связи между авторами (отношения сотрудничества) изображаются фиолетовыми ребрами, а связь авторов с темами исследований показана ребрами серого цвета. В визуализации отображаются темы, активные в данный момент времени. Уровень активности по каждой теме изображается цветом соответствующей вершины, наиболее активные темы в данный момент времени выделяются красным цветом.

На Рис. 14 показано, как изменялись научные интересы и соавторы у автора по имени Жан-Даниэль Фекет (Jean-Daniel Fekete). В период с 2002 по 2004 годы Жан-Даниэль Фекет сотрудничал с двумя авторами по теме 4 (graph, layout, tree). В 2004–2006 годах Фекете продолжал сотрудничать с теми же двумя авторами, а также начал сотрудничать с другим автором по теме 18 (network, social, structure). Темы 4 и 18 похожи, это видно и по тематическим словам («граф» и «сеть»), и по местоположению соответствующих узлов в визуализации. Позже в 2011–2013 годах Фекет сотрудничал с новой группой авторов по теме 19 (analysis, display, knowledge). Это другая тема, и она располагается дальше от первых двух тем в визуализации. Очевидно, что такая визуализация более информативна, чем обычные визуализации сетей соавторства.



[2002-2004] [2004-2006] [2011-2013]

Рис. 14. Визуализация отношений сотрудничества одной и той же персоны на протяжении трех разных временных интервалов [32]

В работах CiteRivers [33] и Cite2vec [26] представлены два разных подхода к повышению информативности визуализации обычных сетей цитирования за счет применения средств анализа текстов. В [33] совместно анализируются и визуализируются два различных аспекта множества научных публикаций: тематическая структура статей, включая динамику во времени, и сети цитирования. В ка-

честве основного инструмента анализа этих двух аспектов используется спектральная кластеризация. Затем строится визуализация, представляющая собой гладкую комбинацию представления эволюции тем в коллекции документов в виде графа течений (streamgraph) с кластеризованным графом цитирований документов, представленном в виде дерева потоков (flowgraph) (Рис. 15). Так же, как и в системах Tiara и ThemeRiver, граф течений изображает кластеры документов, выделенные на основе текстового сходства. Для иерархического объединения документов, а также журналов и конференций используется спектральная кластеризация, которую пользователь может настраивать интерактивно. Помимо текстового сходства, кластеризация может осуществляться на основе метаданных публикаций, что позволяет сгруппировать публикации по конференции, на которой они были представлены, или же по месту работы авторов. При просмотре каждого слоя, соответствующего одному кластеру, можно видеть отдельные временные интервалы. Наиболее частые слова, характеризующие каждый временной интервал, изображаются в виде облака слов. Граф течений, расположенный в левой части изображения, гладко объединяется с изображением потока цитирований, расположенным в правой части изображения. Это соответствует метафоре потока знаний, который течет от цитируемых сообществ к цитирующим документам.

В нижней части изображения расположена еще одна панель, которая называется панелью агрегации цитирований. Она показывает две кривые, голубую и серую. Голубая кривая показывает средний возраст цитирований для каждого временного интервала блока в выделенном слое. По оси ординат отображается среднее время всех публикаций блока в момент публикации статей этого блока. Эта кривая позволяет понять, насколько далеко назад ссылаются публикации данного блока, а также насколько «модной» является тематика данного блока, поскольку менее модные темы будут потенциально ссылаться на более старые публикации.

Вторая, серая кривая изображает энтропию цитирований вдоль высвеченного потока документов. Энтропия цитирований измеряет разброс публикаций в потоке на основе цитируемых мест публикаций. Это позволяет пользователям оценить, насколько широко документы потока цитируют публикации из разных

академических дисциплин и как это эволюционирует во времени. Энтропия возрастает, например, если документы потока начинают цитировать публикации из нового научного сообщества в дополнение к традиционно цитируемым областям. Визуализация затем обогащается дополнительными вычисленными показателями, такими, как свежесть и новизна идей, изменение влиятельности авторов с течением времени и др.

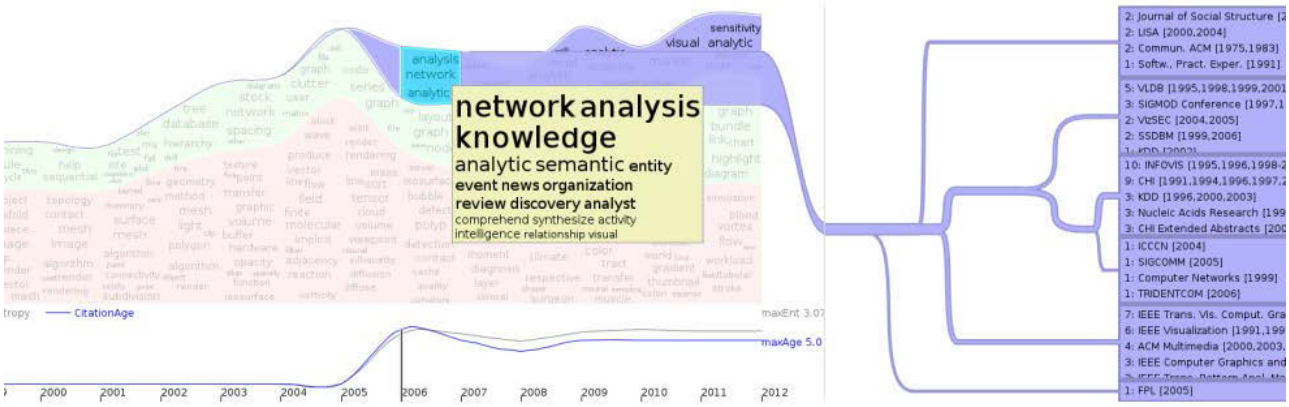
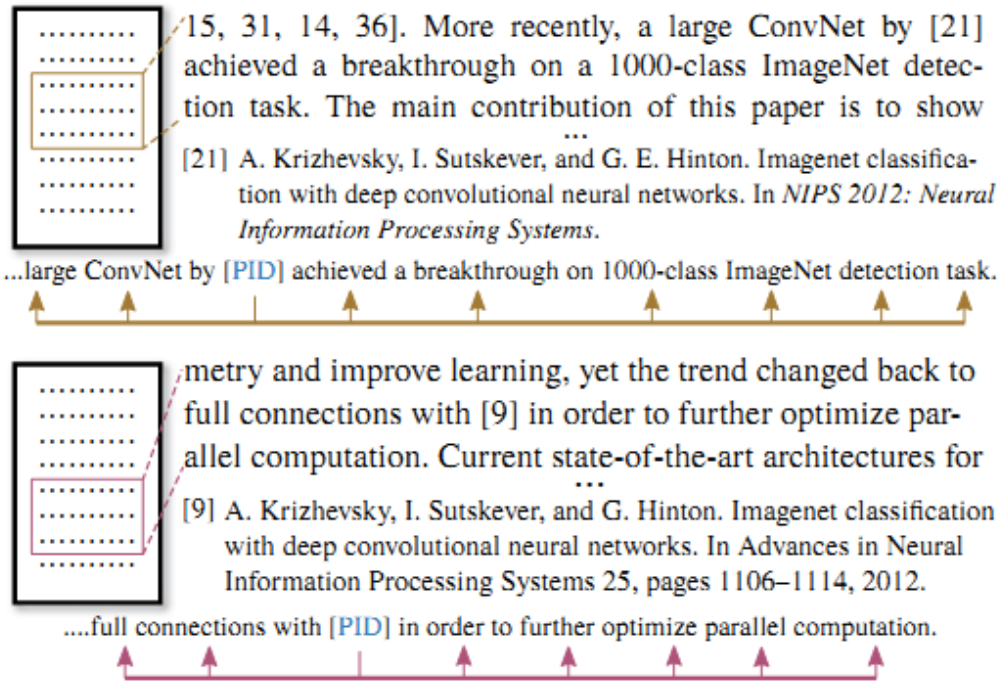


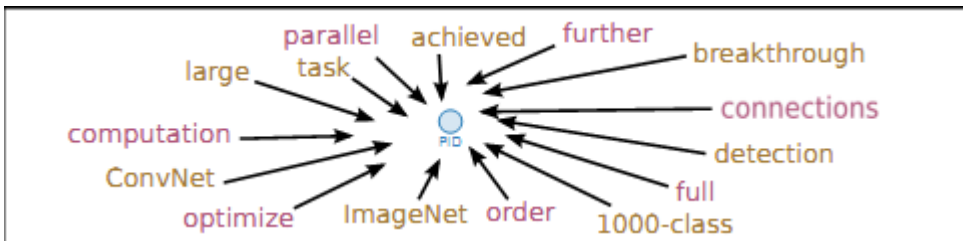
Рис. 15. Совместное представление информации о темах, встречающихся в коллекции документов с информацией о цитированиях документов коллекции в программе CiteRivers[33]

Что касается работы Cite2vec [26], то она применяет для представления коллекции документов модель Skip-gram из набора моделей word2vec. Для обучения векторных моделей используются контексты цитирования научных публикаций, то есть части публикаций, в которых говорится о публикациях, связанных с тематикой конкретной работы (Рис. 16 (а)). Эти части публикаций традиционно носят название «связанные работы». Каждый документ, упоминаемый в списках цитируемой литературы, получает в системе Cite2vec уникальный идентификатор (PIDi). Таким образом, каждому документу сопоставляется уникальное слово, что позволяет представлять и слова, и документы в едином векторном пространстве. При этом вектора, соответствующие словам-идентификаторам доку-

ментов, оказываются в окружении векторов-слов, используемых авторами цитирующих публикаций для описания этих документов.



(а)



(б)

Рис. 16. Подход Cite2vec [26]. (а) На документ, обозначенный PID, ссылаются два разных документа. Контексты цитирования выделены в обоих документах. (б) Контекстные слова расположены рядом с документом PID в пространстве вложения

Для построения визуального представления векторные представления слов, полученные при помощи модели Skip-gram, проецируются на плоскость. Помимо этого, пользователь имеет возможность сформулировать любое понятие, состоящее из нескольких слов. При выборе такого понятия все слова, входящие в понятие, суммируются со словами, присутствующими в визуализации, и исследуемый документ перемещается ближе к тому слову, чья сумма со слова-

ми выбранного понятия наиболее точно отражает смысл рассматриваемого документа. Система позволяет пользователю динамически просматривать документы на основе информации о том, как остальные документы их используют. Например, кто-то ищет публикации про бенчмарки, кто-то про наборы данных и т. д. Пользователю позволено в интерактивном режиме исследовать коллекцию при помощи произвольных словесных фраз, а не при помощи фиксированных наборов слов, связанных с темами.

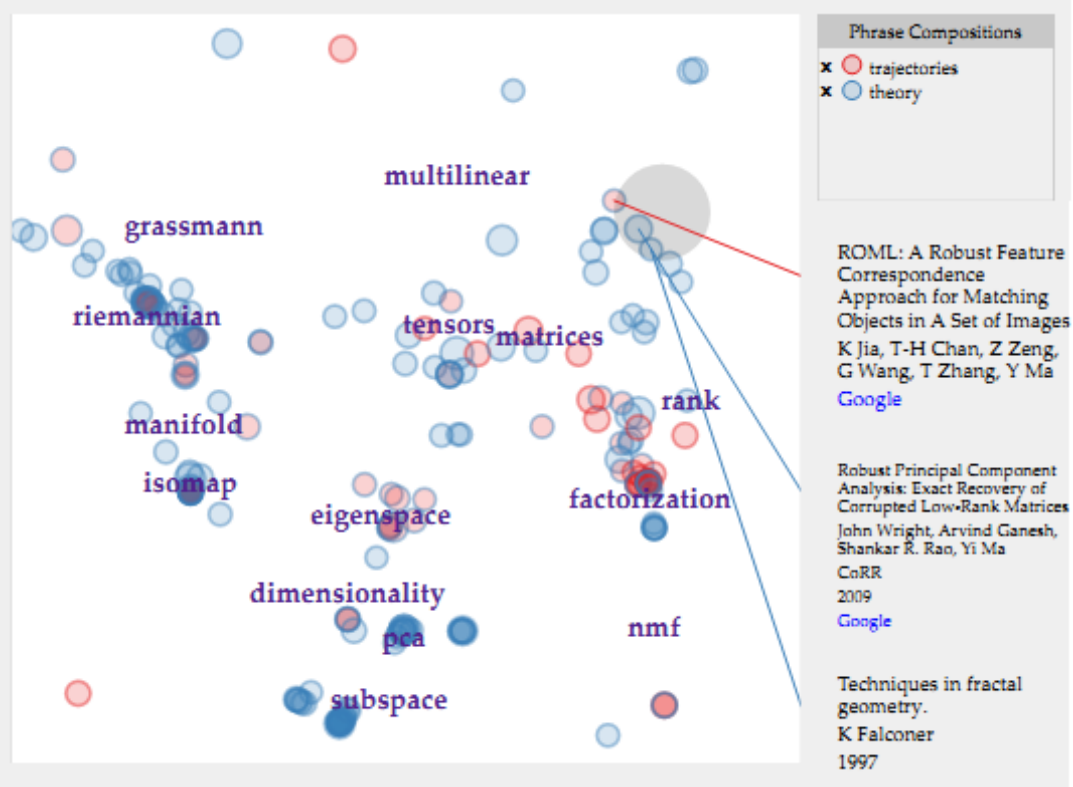


Рис. 17. Показано, как в программе Cite2vec [26] фраза «теория» позволяет находить статьи с теоретическим содержанием

На Рис. 17 показаны преимущества задания нескольких фраз для исследования документов. В данном случае ищутся публикации, посвященные обработке траекторий, а также статьи, которые считаются теоретическими. На рисунке можно видеть, что программа cite2vec обнаруживает статью (ROML), которая ориентирована на последовательности изображений на основе траекторий, а также использует понятия моделирования низкого ранга и предоставление теоретических оценок их подхода. Такие понятия, как «теория», трудно представить в подходе на основе документов, так как язык теоретических статей использует разные термины, которые изменяются в зависимости от области исследований.

Однако подход, основанный на цитировании, может легко справиться с этим, поскольку при цитировании теоретической статьи авторы склонны указывать на характер исследовательского вклада в цитируемой статье.

8. ГИБКАЯ ИНТЕГРАЦИЯ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА И ВИЗУАЛИЗАЦИИ

Часто задачи, с которыми работают исследователи, не сводятся к простому поиску нужного документа. Например, достаточно типичной является ситуация, когда научный сотрудник знакомится с новой областью исследований и стремится понять ключевые идеи, темы и тенденции новой области, а также получить информацию, касающуюся ведущих исследователей, их интересов и сотрудничества. В такой ситуации необходимы просмотр и исследование основных тем, понятий и сущностей, встречающихся в документах, а также понимание связей и отношений между сущностями. Эта проблема не может быть решена средствами одного только интеллектуального анализа данных, поскольку всегда могут возникнуть вопросы, не предусмотренные текущим набором алгоритмов анализа, или результаты недостаточно точны, чтобы сделать вывод. С другой стороны, одна только интерактивная визуализация может тоже оказаться недостаточной для восприятия: по мере того, как растет размер коллекции документов, интерактивное исследование отдельных характеристик каждого документа может просто занять слишком много времени. Для работы с такими ситуациями разработан подход, который называется *визуальная аналитика*, состоящий в гибкой интеграции интеллектуального анализа данных с визуализацией. В [34] представлена версия Jigsaw, предназначенная, в частности, для анализа научной литературы. Система способна помочь новичку разобраться в новой предметной области, понять основную тематику, тенденции, выявить наиболее значительных исследователей, найти ответы на такие вопросы, как:

— Каковы основные темы исследовательских областей разных конференций?

— Как эти темы изменяются со временем?

— Кто является заметным исследователем?

— Какие исследователи в каких областях специализируются?

— Как найти специфические статьи, имеющие отношение к моей текущей области интересов?

При работе с другими предметными областями система способа помочь получить информацию о каком-то заболевании, которым страдает член семьи, или разобраться с огромным набором «профессиональных обзоров», касающихся товара, который надо приобрести.

В системе реализованы три модуля аналитики: резюмирование документов, определение сходства документов и кластеризация. Имеется также модуль анализа тональности текста. Результаты анализа представляются пользователю в виде нескольких взаимосвязанных представлений (View). Элементы, выбранные пользователем в одном из представлений, отображаются и во всех остальных представлениях.

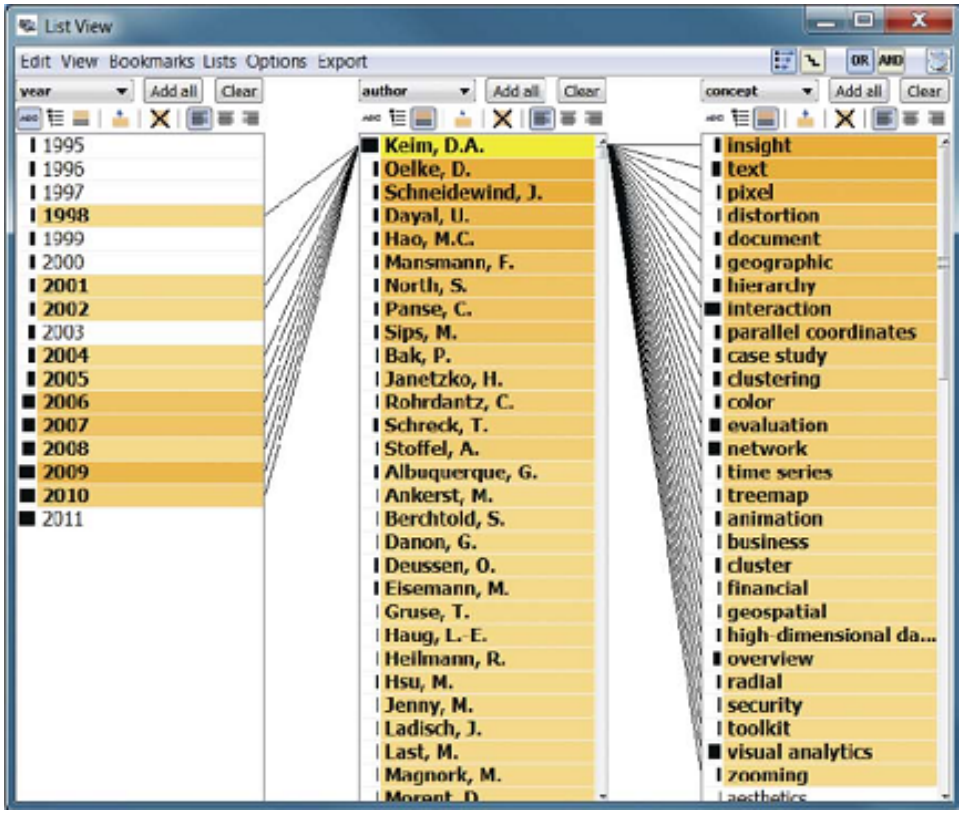


Рис. 18. Визуализация списков сущностей и их взаимосвязей в системе Jigsaw [34]

Одним из основных представлений является представление взаимосвязанных списков сущностей (List View). Сущности, изображаемые в этих списках, могут быть как элементами метаданных публикаций (например, год или место публикации) так и понятиями, извлеченными из текстов публикаций, например,

основными терминами, встречающимися в тексте документа. На Рис. 18 показана визуализация списков сущностей и их взаимосвязей. В центральном списке выбран один из авторов, показанный желтым цветом (Keim, D.A.), что позволяет увидеть в левом списке годы, на которые приходятся публикации этого автора, а в правом списке – понятия, которым посвящены его публикации. Размер небольшого черного прямоугольника слева от каждого элемента списка пропорционален значимости элемента, вычисляемой как количество публикаций, связанных с данным элементом списка.

Рис. 19. Визуализация кластеров в системе Jigsaw [34]

Чтобы лучше понять темы различных конференций на основе названий статей и аннотаций, исследователь может переключиться на представление кластеров документов (Document Cluster View). В этом представлении (Рис. 19, справа) показано 578 документов, разделенных на 20 кластеров, полученных при помощи кластерного анализа. Каждой группе присвоены разные цвета, и каждый кластер помечен тремя описательными ключевыми словами, обычно встречающимися в заголовках и аннотациях в каждом кластере. Если бы термины аннотаций выбирались исключительно на основе их частоты, общие термины, такие, как «данные» и «визуализация», соответствовали бы каждому из кла-

стеров, что не является очень полезным. Поэтому в представлении кластеров документов имеется ползунок частоты, с помощью которого можно выделять или более общие, или более уникальные термины, связанные с каждым кластером.

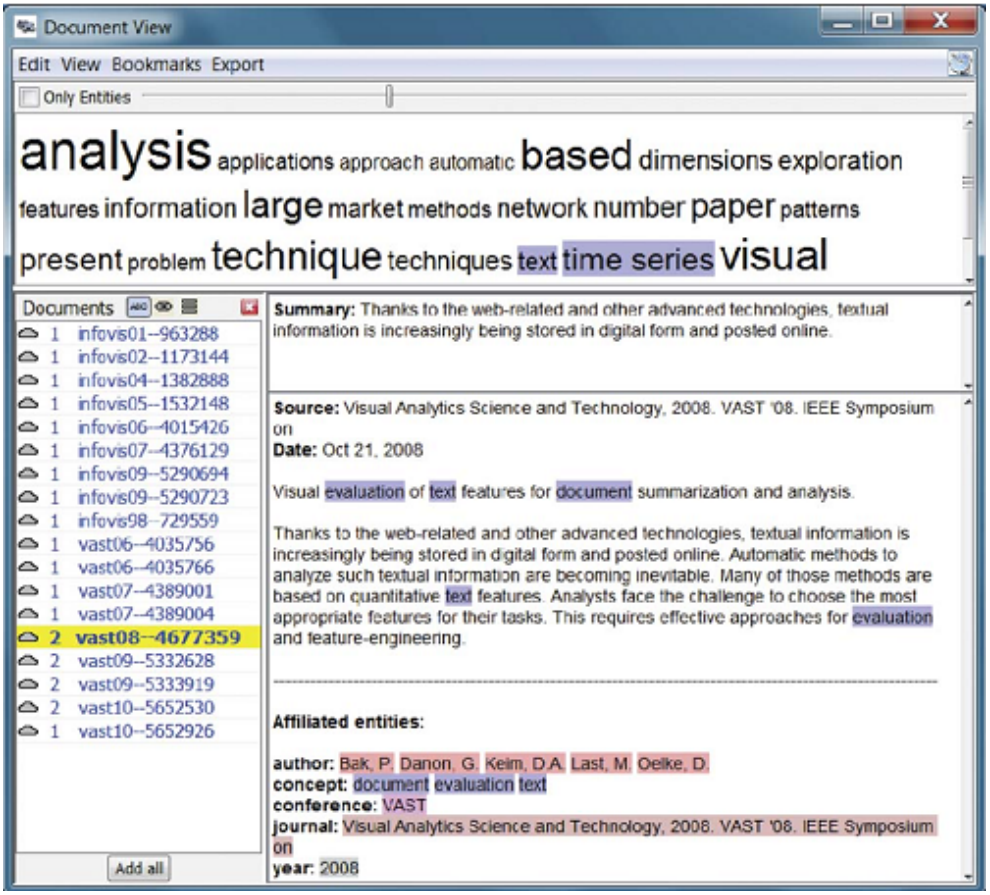


Рис. 20. Визуализация отдельного документа в системе Jigsaw [34]

Поскольку в Представлении Списка документов выбран один элемент (Keim), то все статьи Кейма в Представлении Кластеров Документов выделяются желтыми кругами вокруг прямоугольников документа. Система Jigsaw позволяет выбрать произвольный кластер и исследовать подробнее документы, попадающие в кластер, при помощи представления документа (Document View), показанного на Рис. 20. Слева расположен список документов выбранного кластера, желтым цветом показан один выбранный документ, а справа представлена информация об этом документе. Облако слов наверху показывает наиболее распространенные слова (с высвеченными ключевыми словами и понятиями) в аннотациях загруженных статей. Ниже текста находятся ассоциированные сущности, а над текстом – резюме документа в одной фразе. Такое представление

позволяет быстро понять, соответствуют ли содержание выделенной статьи тому, что ищет исследователь.

Интеграция методов анализа коллекции документов и интерактивной визуализации осуществляется в соответствии со следующими принципами:

1. Разные результаты вычислительного анализа должны быть доступны в любой точке системы в разных контекстах и представлениях, а не в единственном каноническом представлении;

2. Результаты оценки сходства, кластеризации и анализа тональности текста представляются визуально, но данные (множество документов или отдельный документ) могут быть выбраны для последующего анализа;

3. Имеется возможность комбинирования разных значений, получаемых при разных измерениях в одной визуализации;

4. Система должна позволять двигаться в обоих направлениях: как сужая, так и расширяя рамки исследования. Кластеризация документов и анализ тональности сужают рамки до одного кластера или одного документа, а сходство документов и предложение связанных сущностей расширяет рамки;

5. Необходимо предоставлять интерактивный доступ к параметрам алгоритмов.

ЗАКЛЮЧЕНИЕ

В данной работе представлены последние достижения в области визуализации больших текстовых коллекций, совмещающие в себе методы визуализации текстовой информации и метаданных и обладающие высокой интерактивностью при взаимодействии с пользователем. Для современного этапа визуализации коллекций научных документов характерны следующие тенденции:

— комплексный подход, основанный на нескольких алгоритмах анализа и взаимосвязанных интерактивных визуализациях;

— совместное использование методов анализа текста и метаданных для визуализации;

— тесная интеграция множественных представлений;

— интерактивный характер визуализации (визуализация изменяется в результате взаимодействия с пользователем).

СПИСОК ЛИТЕРАТУРЫ

1. *Garfield E.* Historiographic Mapping of Knowledge Domains Literature// J. Inform. Sci. 2004. V. 30, No. 2. P. 119–145.

2. *Apanovich Z.V.* Problems of Visualization of Citation Networks for Large-Science-Portals //ROMAI J. 2012. V. 8, No. 2. P. 13–26.

3. *Small H.* Visualizing Science by Citation Mapping// J. Amer. Soc. Inform. Sci. 1999. V. 50, No. 9. P. 799–813.

4. *Henry N., Fekete J.-D., Mcguffin M.* Nodetrix: A Hybrid Visualization of Social Networks // IEEE Trans. Vis. Comput. Graphics. 2007. V. 13, No. 6. P. 1302–1309.

5. *Gan Q., Zhu M., Li M., Liang T., Cao Y., Zhou B.* Document Visualization: An Overview of Current Research// Wiley Interdisciplinary Reviews: Computational Statistics. 2014. V. 6, No. 1. P. 19–36.

6. *Apanovich Z.V., Vinokurov P.S., Elagin V. A.* An Approach to visualization of knowledge portal content // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2009. T. 29. P. 17-32.

7. *Strobelt H, Oelke D, Rohrdantz C, Stoffel A, Keim D.A., Deussen O.* Document Cards: A Top Trumps Visualization for Dcuments// IEEE Trans Vis Comput Graph. 2009. V. 15. P. 1145–1152.

8. *Schulz H.-J.* Treevis.net: A Tree Visualization Reference // IEEE Computer Graphics and Applications. 2011. V. 31, No. 6. P. 11–15.

9. *Aigner W., Miksch S., Schumann H., Tominski C.* Visualization of Time-Oriented Data. Springer, 2011. 286 p.

10. *Kucher K., Kerren A.* Text Visualization Techniques: Taxonomy, Visual Survey, and Community Insights // Proc. of the 8th IEEE PacificVis. 2015. P. 117–121.

11. *Beck F., Koch S., Weiskopf D.* Visual analysis and Dissemination of Scientific Literature Collections with SurVis // IEEE Trans.Vis. Comput. Graphics. 2016. V. 22, No. 1. P. 180–189.

12. *Hofmann T.* Probabilistic Latent Semantic Indexing // Proc. the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). 1999. P. 50–57.

13. *Blei D.M.* Probabilistic Topic Models// Communications of the ACM. 2012. V. 55, No. 4. P. 77–84.

14. *Alexander E., Kohlmann J., Valenza R., Gleicher M.* Serendip: Turning Topics Back to the Text // 2013 IEEE Visualization Poster Proc. (InfoVis ’13).

15. *Chuang J., Manning C.D., Heer J.* 2012. Termite: Visualization Techniques for Assessing Textual Topic Models// Proc. of the Int. Working Conf. on Advanced Visual Interfaces. ACM, 2012. P. 74–77.

16. *Chaney A.J.-B., Blei D.M.* Visualizing Topic Models // Int. AAAI Conf. on Social Media and Weblogs, 2012. P. 419–422.

17. L*ee H., Kihm J., Choo J., Stasko J., Park H.* iVisClustering: An Interactive Visual Document Clustering Via Topic Modeling// Computer Graphics Forum (CGF). 2012. V. 31. P. 1155–1164.

18. *Dou W., Wang X., Chang R., Ribarsky W.* ParallelTopics: A Probabilistic Approach to Exploring Document Collections // Proc. of IEEE Conf. on Visual Analytics Science and Technology. 2011. P. 231–240.

19. *Blundell C., Teh Y.W., Heller K.A.* Bayesian Rose Trees // Proc. Int. Conf. Uncertainty Artif. Intell. 2010. P. 65–72.

20. *Liu S., Wang X., Chen J., Zhu J., Guo B.* TopicPanorama: A Full Picture of Relevant Topics // Proc. of the IEEE Conf. on Visual Analytics Science and Technology (VAST). 2014. P. 183–192.

21. *Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei.* How Hierarchical Topics Evolve in Large Text Corpora // IEEE Trans. Vis. Comput. Graph. 2014. V. 20, No. 12. P. 2281–2290.

22. *Kuang D., H. Park.* Fast Rank-2 Nonnegative Matrix Factorization for Hierarchical Document Clustering // Proc. the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). 2013. P. 739–747.

23. *Choo J., Lee C., Reddy C.K., Park H.* Utopian: User-driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization // IEEE Transactions on Visualization and Computer Graphics. 2013. V. 19, No. 12. P. 1992–2001.

24. *Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, Niklas Elmqvist.* TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections// IEEE Transactions on Visualization and Computer Graphics. 2017. V. 23, No. 1. P. 151–160.

25. *Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J.* Distributed Rrepresentations of Words and Phrases and Their Compositionality// Advances in Neural Information Processing Systems. 2013. P. 3111–3119.

26. *Berger M., McDonough K., Seversky Lee M.* Cite2vec: Citation-Driven Document Exploration via Word Embeddings // IEEE Transactions on Visualization and Computer Graphics. 2017. V. 23, No. 1. P. 691–700.

27. *Steyvers M., Griffiths T.* Probabilistic Topic Models // Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum. 2007. P. 1–15.

28. *Havre S., Hetzler E., Whitney P., Nowel L.l.* ThemeRiver: Visualizing Thematic Changes in Large Document Collections // IEEE Transactions on Visualization and Computer Graphics (TVCG). 2002. V. 8, No. 1. P. 9–20.

29. *Wei F., Liu S., Song Y., Pan S., Zhou M.X., Qian W., Shi L., Tan L., Zhang Q.* TIARA: A Visual Exploratory Text Analytic System // Proc. the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). 2010. P. 153–162.

30. *Sugiyama K., Tagawa S., Toda M.* Methods for Visual Understanding of Hierarchical System Structures // IEEE Transactions on Systems, Man and Cybernetics. 1981. V. 11, No. 2. P. 109–125.

31. *Gretarsson B., O'Donovan J., Bostandjiev S., H¨ ollerer T., Asuncion A.,* Newman *D.*, *Smyth P.* TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling // ACM Transactions on Intelligent Systems and Technology (TIST). 2012. V. 3, No. 2. P. 1–26.

32. *Chen F., Chiu P., Lim S.* Topic Modeling of Document Metadata for Visualizing Collaborations over Time // Proc. of the Int. Conf. on Intelligent User Interfaces (IUI). 2016. P. 108–117.

33. *Heimerl F., Qi Han, Koch S., Ertl T.* CiteRivers: Visual Analytics of Citation Patterns // IEEE Transactions on Visualization and Computer Graphics 1. 2016. V. 22, No. 1. P. 190–199.

34. *Görg C., Liu Zh., Kihm J., Ch Jaegul, Park H., Stasko J.* Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw// IEEE Transactions on Visualization and Computer Graphics. 2013. V. 19, No. 10. P. 1646–1663.

EVOLUTION OF VISUALIZATION METHODS FOR RESEARCH PUBLICATION COLLECTIONS

Z. V. Apanovich

A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk State University, Novosibirsk

apanovich@iis.nsk.su

Abstract

The information visualization methods have been known as a tool providing the understanding of large data. The visualization of research publication collections is a special case of applying visualization methods to large data. This paper presents a survey of problems solved by means of visualization, document models and document analysis methods as well as of new approaches to visualization methods for research publication collections. Special attention is paid to the relation between the document analysis and visualization methods.

Keywords*: visualization of document collections, text analysis, text and metadata visualization algorithms, LDA, NMF, word2vec*

REFERENCES

1. *Garfield E.* Historiographic Mapping of Knowledge Domains Literature// J. Inform. Sci. 2004. V. 30, No. 2. P. 119–145.

2. *Apanovich Z.V.* Problems of Visualization of Citation Networks for Large-Science-Portals //ROMAI J. 2012. V. 8, No. 2. P. 13–26.

3. *Small H.* Visualizing Science by Citation Mapping// J. Amer. Soc. Inform. Sci. 1999. V. 50, No. 9. P. 799–813.

4. *Henry N., Fekete J.-D., Mcguffin M.* Nodetrix: A Hybrid Visualization of Social Networks // IEEE Trans. Vis. Comput. Graphics. 2007. V. 13, No. 6. P. 1302–1309.

5. *Gan Q., Zhu M., Li M., Liang T., Cao Y., Zhou B.* Document Visualization: An Overview of Current Research// Wiley Interdisciplinary Reviews: Computational Statistics. 2014. V. 6, No. 1. P. 19–36.

6. *Apanovich Z.V., Vinokurov P.S., Elagin V. A.* An Approach to visualization of knowledge portal content // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2009. T. 29. P. 17–32.

7. *Strobelt H, Oelke D, Rohrdantz C, Stoffel A, Keim D.A., Deussen O.* Document Cards: A Top Trumps Visualization for Dcuments// IEEE Trans Vis Comput Graph. 2009. V. 15. P. 1145–1152.

8. *Schulz H.-J.* Treevis.net: A Tree Visualization Reference // IEEE Computer Graphics and Applications. 2011. V. 31, No. 6. P. 11–15.

9. *Aigner W., Miksch S., Schumann H., Tominski C.* Visualization of Time-Oriented Data. Springer, 2011. 286 p.

10. *Kucher K., Kerren A.* Text Visualization Techniques: Taxonomy, Visual Survey, and Community Insights // Proc. of the 8th IEEE PacificVis. 2015. P. 117–121.

11. *Beck F., Koch S., Weiskopf D.* Visual analysis and Dissemination of Scientific Literature Collections with SurVis // IEEE Trans.Vis. Comput. Graphics. 2016. V. 22, No. 1. P. 180–189.

12. *Hofmann T.* Probabilistic Latent Semantic Indexing // Proc. the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). 1999. P. 50–57.

13. *Blei D.M.* Probabilistic Topic Models// Communications of the ACM. 2012. V. 55, No. 4. P. 77–84.

14. *Alexander E., Kohlmann J., Valenza R., Gleicher M.* Serendip: Turning Topics Back to the Text // 2013 IEEE Visualization Poster Proc. (InfoVis ’13).

15. *Chuang J., Manning C.D., Heer J.* 2012. Termite: Visualization Techniques for Assessing Textual Topic Models// Proc. of the Int. Working Conf. on Advanced Visual Interfaces. ACM, 2012. P. 74–77.

16. *Chaney A.J.-B., Blei D.M.* Visualizing Topic Models // Int. AAAI Conf. on Social Media and Weblogs, 2012. P. 419–422.

17. L*ee H., Kihm J., Choo J., Stasko J., Park H.* iVisClustering: An Interactive Visual Document Clustering Via Topic Modeling// Computer Graphics Forum (CGF). 2012. V. 31. P. 1155–1164.

18. *Dou W., Wang X., Chang R., Ribarsky W.* ParallelTopics: A Probabilistic Approach to Exploring Document Collections // Proc. of IEEE Conf. on Visual Analytics Science and Technology. 2011. P. 231–240.

19. *Blundell C., Teh Y.W., Heller K.A.* Bayesian Rose Trees // Proc. Int. Conf. Uncertainty Artif. Intell. 2010. P. 65–72.

20. *Liu S., Wang X., Chen J., Zhu J., Guo B.* TopicPanorama: A Full Picture of Relevant Topics // Proc. of the IEEE Conf. on Visual Analytics Science and Technology (VAST). 2014. P. 183–192.

21. *Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei.* How Hierarchical Topics Evolve in Large Text Corpora // IEEE Trans. Vis. Comput. Graph. 2014. V. 20, No. 12. P. 2281–2290.

22. *Kuang D., H. Park.* Fast Rank-2 Nonnegative Matrix Factorization for Hierarchical Document Clustering // Proc. the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). 2013. P. 739–747.

23. *Choo J., Lee C., Reddy C.K., Park H.* Utopian: User-driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization // IEEE Transactions on Visualization and Computer Graphics. 2013. V. 19, No. 12. P. 1992–2001.

24. *Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, Niklas Elmqvist.* TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections// IEEE Transactions on Visualization and Computer Graphics. 2017. V. 23, No. 1. P. 151–160.

25. *Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J.* Distributed Rrepresentations of Words and Phrases and Their Compositionality// Advances in Neural Information Processing Systems. 2013. P. 3111–3119.

26. *Berger M., McDonough K., Seversky Lee M.* Cite2vec: Citation-Driven Document Exploration via Word Embeddings // IEEE Transactions on Visualization and Computer Graphics. 2017. V. 23, No. 1. P. 691–700.

27. *Steyvers M., Griffiths T.* Probabilistic Topic Models // Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum. 2007. P. 1–15.

28. *Havre S., Hetzler E., Whitney P., Nowel L.l.* ThemeRiver: Visualizing Thematic Changes in Large Document Collections // IEEE Transactions on Visualization and Computer Graphics (TVCG). 2002. V. 8, No. 1. P. 9–20.

29. *Wei F., Liu S., Song Y., Pan S., Zhou M.X., Qian W., Shi L., Tan L., Zhang Q.* TIARA: A Visual Exploratory Text Analytic System // Proc. the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). 2010. P. 153–162.

30. *Sugiyama K., Tagawa S., Toda M.* Methods for Visual Understanding of Hierarchical System Structures // IEEE Transactions on Systems, Man and Cybernetics. 1981. V. 11, No. 2. P. 109–125.

31. *Gretarsson B., O'Donovan J., Bostandjiev S., H¨ ollerer T., Asuncion A.,* Newman *D.*, *Smyth P.* TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling // ACM Transactions on Intelligent Systems and Technology (TIST). 2012. V. 3, No. 2. P. 1–26.

32. *Chen F., Chiu P., Lim S.* Topic Modeling of Document Metadata for Visualizing Collaborations over Time // Proc. of the Int. Conf. on Intelligent User Interfaces (IUI). 2016. P. 108–117.

33. *Heimerl F., Qi Han, Koch S., Ertl T.* CiteRivers: Visual Analytics of Citation Patterns // IEEE Transactions on Visualization and Computer Graphics 1. 2016. V. 22, No. 1. P. 190–199.

34. *Görg C., Liu Zh., Kihm J., Ch Jaegul, Park H., Stasko J.* Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw// IEEE Transactions on Visualization and Computer Graphics. 2013. V. 19, No. 10. P. 1646–1663.

СВЕДЕНИЯ ОБ АВТОРЕ

АПАНОВИЧ Зинаида Владимировна – старший научный сотрудник Института Систем Информатики СО РАН, доцент Новосибирского государственного университета. Сфера научных интересов – визуализация информации, визуализация графов, Semantic Web

APANOVICH Zinaida Vladimirovna – senior researcher of the Institute of Informatics Systems of SB RAS, Associate Professor of Novosibirsk State University. Research interests include information visualization, graph visualization, Semantic Web.

email: apanovich@iis.nsk.su

Материал поступил в редакцию 11 сентября 2017 года