ИПМ им.М.В.Келдыша РАН

Абрау-2017 • Труды конференции

З.В. Апанович

Современные тенденции визуализации коллекций научных публикаций

Рекомендуемая форма библиографической ссылки

Апанович З.В. Современные тенденции визуализации коллекций научных публикаций // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2017. — С. 3-8. — URL: http://keldysh.ru/abrau/2017/36.pdf doi:10.20948/abrau-2017-36

Размещена также презентация к докладу

Современные тенденции визуализации коллекций научных публикаций

З.В. Апанович

Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск
apanovich@iis.nsk.su

Аннотация. По мере возрастания количества публикаций, доступных в Интернете, возрастает необходимость в инструментах, позволяющих ориентироваться в огромном потоке информации. Одним из инструментов, стремительно развивающимся в последние годы являются методы визуализации больших коллекций научных публикаций. В статье рассматриваются задачи, решаемые при помощи визуализации, модели представления и методы анализа текстовой информации, а также новые подходы к визуализации документов.

Ключевые слова: визуализация коллекций документов, анализ текстов, алгоритмы визуализации текстов и метаданных

Введение

В последние годы текстовая информация в электронном виде генерируется в огромных количествах. Например, только за 2006 год было опубликовано 1.35 миллиона научных статей, среднегодовой рост количества публикаций составляет 2,5%, и в настоящее время публикуется более 4400 названий в день. Все больше научных электронных коллекций предоставляют полные тексты публикаций. Например, сайт SpringerLink[1] предоставляет 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] и др. Тексты 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.

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

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

[3] https://arxiv.org/

[4] https://socionet.ru/

Визуализация документов — это класс методов визуализации информации, преобразующих текстовую информацию, такую как слова, предложения, документы и их взаимоотношения в визуальную форму, позволяя пользователям лучше понимать текстовые документы и уменьшать нагрузку на пользователей при работе с большими коллекциями текстовых документов. 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 [1-5], то с развитием методов анализа и визуализации текстов появляется все больше публикаций, совместно анализирующих и визуализирующих текстовые данные и метаданные.

1. Модели представления текстовой информации

В основе методов визуализации документов лежат различные модели представления текстов. 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. Примером системы, использующей представление TF-IDF для визуализации коллекции документов, является Jigsaw [5]. Недостатком этой модели является высокая размерность векторного пространства.

Модель, которая существенно снижает размерность векторного представления текста, носит название *вероятностное моделирование тем.* 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* (pLSI) [6] и *скрытое распределение Дирихле* (LDA) [7] — два популярных метода в этой категории. Вероятностные тематические модели осуществляют «мягкую» кластеризацию, позволяя документу или термину относиться сразу к нескольким темам с различными вероятностями. 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. На практике в результате применения метода LDA получается более корректный набор тем. Примером использования представления LDA для визуализации коллекций документов

является работа [8]. В случае больших коллекций документов с большим количеством тем, их принято организовывать иерархически с применением модели *Байесовское розо-дерево* BRT[9], которая использует байесовский алгоритм иерархической кластеризации для построения дерева тем с произвольным коэффициентом ветвления, в котором каждая нелистовая вершина изображает кластер тем. Примерами систем визуализации, использующих иерархическое представление моделей тем являются TopicPanorama и HierarchicalTopics [10-11]. Однако общим недостатком всех вариаций модели LDA является высокие требования к производительности. С недавнего времени, *неотрицательное матричное разложение* (NMF), в котором матрица термин-документ представляется в виде произведения двух матриц с неотрицательными элементами [12] используется в качестве альтернативного подхода к моделированию тем в анализе документов. Примером использования представления NMF для визуализации коллекций документов являются [13-14].

Поскольку модель NMF имеет существенно более высокую скорость работы по сравнению с LDA, на ее основе разработано несколько систем, динамически управляющих процессом моделирования тем. Так, в работе [13] 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 основан на дистрибутивной семантике и векторном представлении слов. *Векторное представление слов* основывается 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 [15]. 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 [16].

Для того, чтобы построить отображение векторного пространства высокой размерности на плоскость, принято использовать такие методы как многомерное шкалирование (MDS) и метод главных компонент (PCA). В последние годы стал очень популярным метод снижения размерности векторного пространства называемый t-sne (t-Distributed Stochastic Neighbor Embedding) [17], поскольку оказался очень эффективным средством визуализации структуры высокоразмерных данных.

2. Задачи и тенденции систем визуализации научных коллекций

При визуализации коллекций документов принято выделять следующие подзадачи:

— визуализация основного контента коллекции документов,
— визуализация тем коллекции документов,
— визуализация отношений между документами, в частности визуализации 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.

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).

3. 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 [18], которая визуализировала ключевой контент документов как смесь изображений и важных терминов, благодаря чему изображение напоминало карты в карточной игре. Основным достижением этой работы была демонстрация важности использования изображений в качестве полноправных метаданных, что привело к большому количеству работ, расширяющих этот подход [19-21]. Наконец в работе [22] представлена система SurVis, позволяющая создавать визуальные обзоры коллекций публикаций. Для каждой публикации создается отдельная «карточка» где в правой части приводятся основные метаданные публикации, такие как авторы, название, абстракт, DOI, репрезентативный рисунок из текста публикации, а в левой части, изображение ключевых слов и имен авторов в виде облака слов.

Основным недостатком всех этих визуализаций является то, что не всегда по указанной ссылке доступен полный текст публикации, а рисунок хоть и является уникальным идентификатором публикации, как правило, достаточно сложно интерпретировать, не прочитав подробное описание того, как надо понимать указанное изображение. Кроме этого, такие визуальные обзоры очень

редко содержат информацию об алгоритмах, использованных для создания визуализации.

4. Примеры современных инструментов визуализации научных коллекций

В работе [6] представлена версия Jigsaw для анализа научной литературы. В программе имеется три разных способа реферирования отдельных документов, в частности, имеется возможность представления документа одной наиболее важной фразой. Документы группируются в соответствии с двумя показателями подобия, один основан на анализе текстового контента, а другой - на основе связей между сущностями, извлеченными из метаданных, такими как конференции, места работы, годы. Кластеры документов визуализируются как диаграммы связей узлов, а также в виде сетки, где документы представляются в виде небольших квадратов, упорядоченных и окрашенных в соответствии с их сходством с выбранным документом. Визуализация обогащается результатами базового анализа настроений (в случае применения к анализу всевозможных ревю).

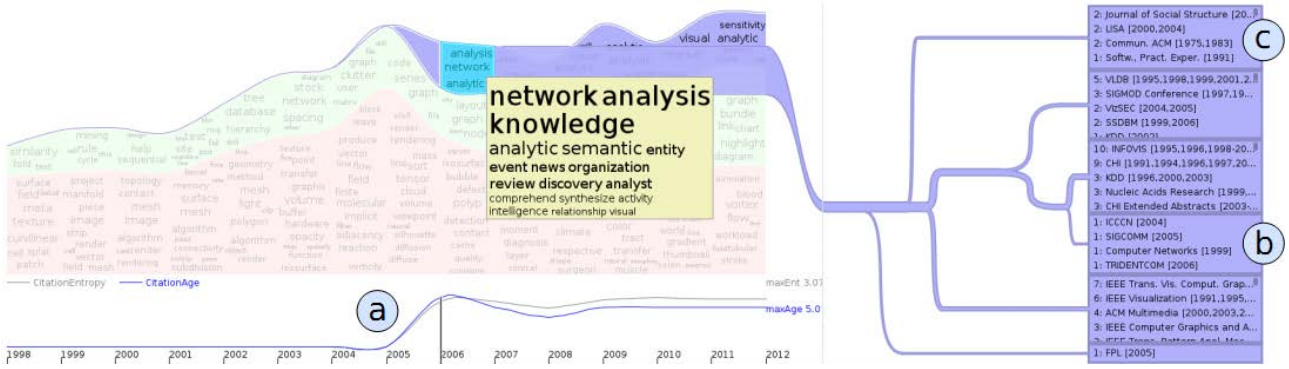

Рис. 1. Совместное представление информации о темах, встречающихся в коллекции документов с информацией о цитированиях документов коллекции в программе CiteRivers[23]

В работе CiteRivers [23] реализована гладкая комбинация представления эволюции тем в коллекции документов в виде графа течений (streamgraph) с кластеризованным графом цитирований документов, представленном в виде дерева потоков (flowgraph) (рис. 1). Для иерархического агрегирования документов, а также журналов и конференций используется спектральная кластеризация, которую пользователь может настраивать интерактивно. Визуализация затем обогащается дополнительными вычисленными показателями, такими как свежесть и новизна идей, изменение влиятельности авторов с течением времени и др.

Подход Cite2vec [16] использует для представления коллекции документов модель Skip-gram из набора моделей word2vec, используя для обучения моделей контексты цитирования научных публикаций (рис. 2).

Каждый документ, упоминаемый в списках научной литературы, представляется уникальным идентификатором (Idi), что позволяет представлять слова и документы в едином векторном пространстве. Система позволяет пользователю динамически просматривать документы на основе информации о том, как остальные документы используют их. 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.

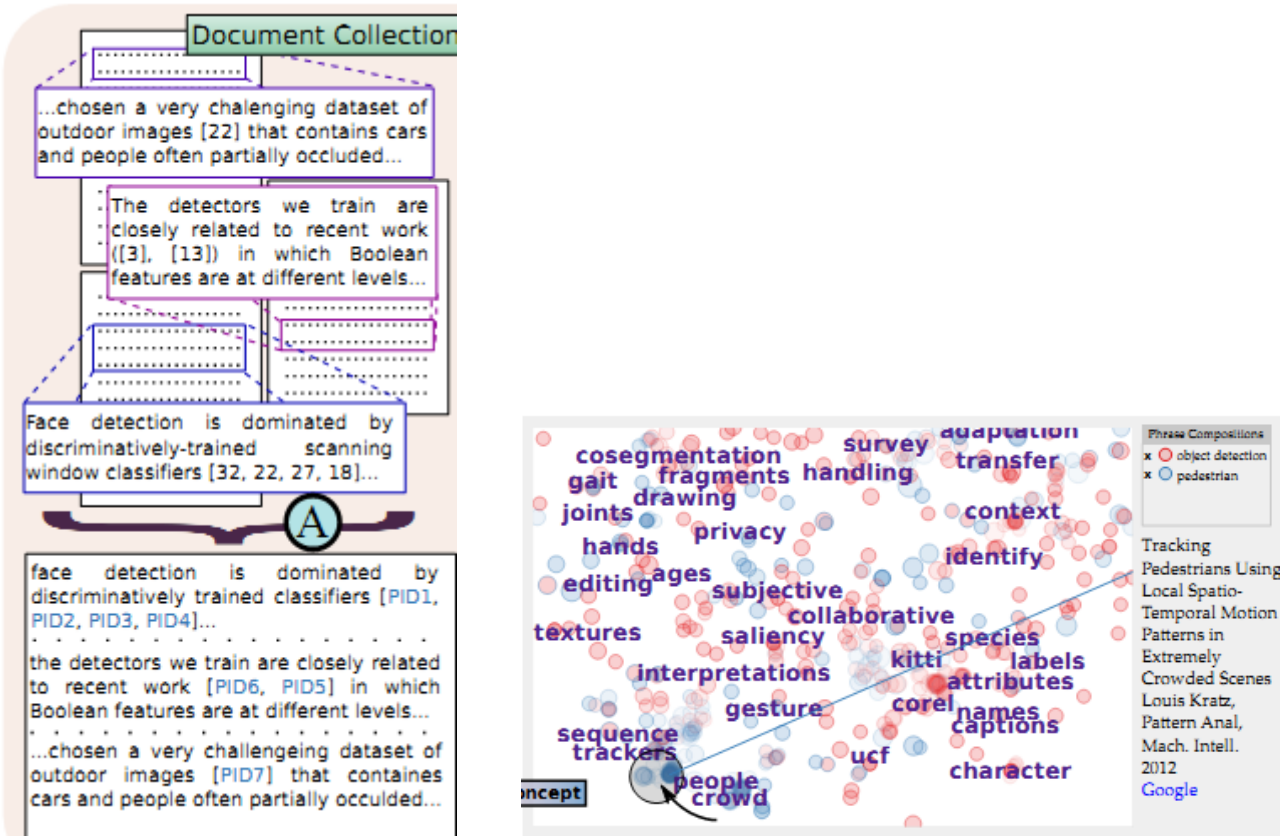

Рис. 2. Представление документа в виде одного слова и совместное изображение слов и документов в едином векторном пространстве в программе Cite2vec.

В работе [24] метод LDA используется для автоматического вычисления областей интереса на основе метаданных научных публикаций, таких как название, авторы, ключевые слова и абстракт. Темы, извлеченные из документов, изображаются в виде фиксированных вершин, а авторы публикаций, работающие над разными темами в разные моменты времени, изображаются перемещаемыми вершинами. Для каждой темы подсчитывается количество документов в каждый момент времени и генерируется поток активности.

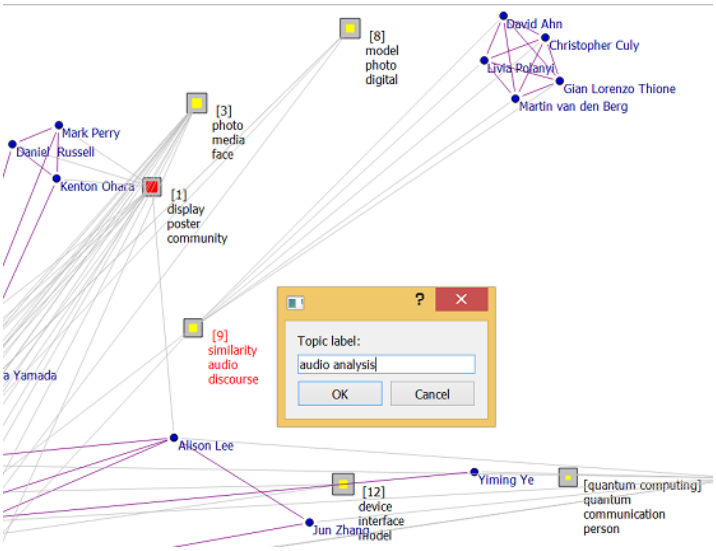

Рис. 3. Показана подробная информация об отношениях соавторства: Связи между авторами, темы над которыми они сотрудничают, уровни всплесков тем за выбранный период времени.

TopicLence [14] является развитием похода, реализованного в [13]. 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. См. рис 4.

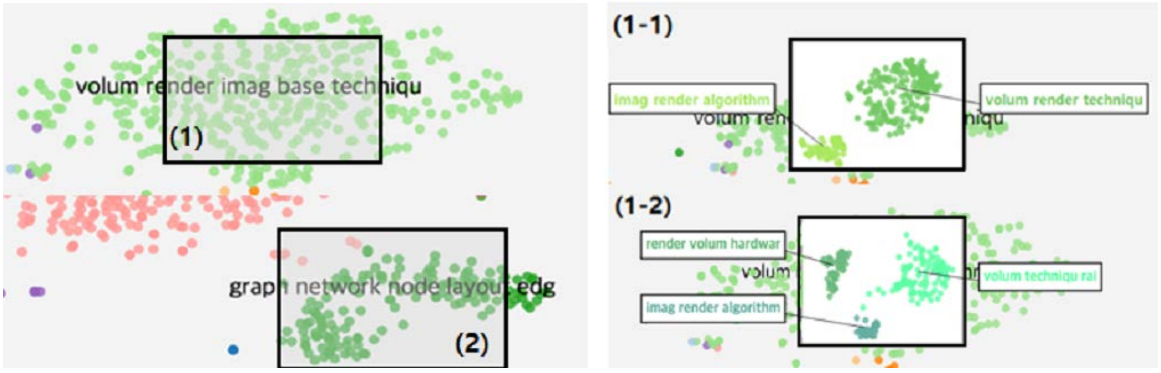

Рис. 4. (а) Начальное моделирование тем, (б) Уточнение тем для области 1

Также следует отметить появление новых применений визуализации текстовых коллекций. Например, данные об авторах по-прежнему страдают от двусмысленностей из-за синонимов (другое правописание, один и то же человек) и омонимов (одно и то же имя, разные лица). В последние годы появляются специальные методы визуализации для поддержки дедупликации имен и устранения неоднозначности [25, 26].

5. Заключение

В данной работе представлены последние достижения в области визуализации больших текстовых коллекций, совмещающие в себе методы визуализации текстовой информации совместно с метаданными и обладающие высокой интерактивностью при взаимодействии с пользователем.

Литература

1. Garfield E., “Historiographic mapping of knowledge domains literature// J. Inform. Sci., . — vol. 30, no. 2. — 2004. — pp. 119–145
2. Z. V. Apanovich, Problems of visualization of citation networks for large-science-portals, // ROMAI Journal. — Vol.8, Nr.2. —2012. — pp. 13-26
3 H. Small, “Visualizing science by citation mapping// J. Amer. Soc. Inform. Sci. — vol. 50, no. 9. —1999. — pp. 799–813
4. N. Henry, J.-D. Fekete, and M. McGuffin,NodeTrix: a hybrid visualization of social networks// IEEE Trans. Vis. Comput. Graphics. — vol. 13, no. 6. — 2007. —pp. 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 6 (1) . — pp. 19-36
6. 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. — vol. 19, no. 10. — 2013. —pp. 1646-1663.
7. Hofmann T. Probabilistic latent semantic indexing//Proc. the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). — 1999. — pp. 50–57.
7. D. M. Blei. Probabilistic topic models // Communications of the ACM. —55(4). — 2012. —pp. 77–84.
8. Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In ICWSM
9. Blundell C., Teh Y. W., Heller K. A., “Bayesian rose trees,” in Proc. Int. Conf. Uncertainty Artif. Intell., 2010. — pp. 65–72.
10. Liu S., Wang X., Chen J., Zhu J., Guo B.. TopicPanorama: a full picture of relevant topics. In Proc. theof the IEEE Conference on Visual Analytics Science and Technology (VAST). — 2014. — pp. 183–192.
11. Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei: How Hierarchical Topics Evolve in Large Text Corpora. IEEE Trans. Vis. Comput. Graph. —. 20(12). — 2014. — pp. 2281-2290
12. Kuang D., H. Park. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering//Proc. the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). — 2013. —pp. 739–747.
13. J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. Visualization and Computer Graphics, IEEE Transactions on. — 19(12) 2013. — pp.1992–2001.

14. Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, and Niklas Elmqvist. TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections // IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1 . — 2017. —pp. 151-160.
15. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., Distributed representations of words and phrases and their compositionality// Advances in neural information processing systems . — 2013. —pp. 3111–3119.
16. Berger M., McDonough K., Seversky Lee M.. cite2vec: Citation-Driven Document Exploration via Word Embeddings// IEEE Transactions on Visualization and Computer Graphics. — vol. 23, no. 1. —2017. — pp. 691-700
17. Van Der Maaten L. Accelerating t-sne using tree-based algorithms //The Journal of Machine Learning Research. — 15(1). — 2014. —pp 3221–3245.
18. Strobelt H, Oelke D, Rohrdantz C, Stoffel A, Keim DA, Deussen O. Document cards: a top trumps visualization for documents// IEEE Trans Vis Comput Graph . —2009. —15. —pp. 1145–1152.
19. Aigner W., Miksch S., Schumann H., Tominski C.. Visualization of Time-Oriented Data // Springer. — 2011.
20. Schulz H.-J., Treevis.net: A tree visualization reference// IEEE Computer Graphics and Applications. — 31(6) . — 2011.— pp. 11–15.
21. K. Kucher and A. Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights // Proceedings of the 8th IEEE PacificVis. — . 2015— pp. 117–121
22. Beck F., Koch S., Weiskopf. Visual analysis and dissemination of scientific literature collections with SurVis, // IEEE Trans.Vis. Comput. Graphics1. — vol. 22, no. 1. — 2016.
23. Heimerl F., Qi Han, Koch S., Ertl T. CiteRivers: Visual Analytics of Citation Patterns. // IEEE Transactions on Visualization and Computer Graphics, 1. — vol. 22, no. 1. — 2016 — pp. 190-199.
24. Chen F., Chiu P., Lim S.. Topic Modeling of Document Metadata for Visualizing Collaborations over Time //Proceedings of the International Conference on Intelligent User Interfaces (IUI). — 2016. — pp. 108-117.
25. Апанович З.В. Сопоставление данных разноязычных ресурсов и кросс-языковая идентификация авторов //Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2016. — С. 36-45.
26. Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu, Weiwei Cui. NameClarifier: A Visual Analytics System for Author Name Disambiguation // IEEE Transactions on Visualization and Computer Graphics. — vol. 23, no. 1. — 2017.— pp. 141-150.