DOI: 10.25586/RNU.V9187.18.06.P.46 УДК 681.3.07

О.В. Золотарев, М.М. Шарнин, А.Г. Мацкевич, С.В. Клименко

СЕМАНТИЧЕСКИЙ ПОДХОД К ВИЗУАЛИЗАЦИИ НАУЧНЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ВЕБ-ГРАФИКИ 3D[1]

В статье описывается семантический подход к визуализации 3D-киберпространства научных работ и их исследований с использованием веб-3D-графики. Наиболее цитируемые и значимые документы в этом киберпространстве отображаются сферами большого размера, а расстояние между документами – пропорционально их смысловому сходству. Предложена новая мера семантического подобия документов, которая определяется максимальной корреляцией между явной и неявной связностью документов. Предложен и внедрен новый индекс контекстного цитирования документов (SCCI), который определяется по максимуму корреляции с индексом научного цитирования (SCI). SCCI может более точно измерять научную значимость документов, находить важные документы и оценивать новые статьи с нулевым SCI. Значимые научные статьи подтверждают друг друга и образуют кластеры в киберпространстве. В результате исследования формируется набор таких кластеров. Предлагаемое киберпространство, реализованное в WebVR и с помощью интерактивной 3D-графики, можно рассматривать как динамичную среду обучения, которая удобна для обнаружения новых значимых статей, идей и тенденций.

Ключевые слова*: *киберпространство, веб-3D-графика, WebVR, научные статьи, семантическое сходство, индекс контекстного научного цитирования, неявные ссылки, интеграция информации, визуализация.

O.V. Zolotarev, M.M. Charnine, A.G. Matskevich, S.V. Klimenko

SEMANTIC APPROACH TO VISUALIZATION OF SCIENTIFIC DOCUMENTS USING 3D WEB GRAPHICS

In this paper we describe a semantic approach to visualization of 3D cyberspace of scientific papers and their research front using web-based 3D graphic. The most cited and significant documents in this cyberspace are represented by spheres of a large size, and the distance between documents is proportional to their semantic similarity. A new measure of semantic similarity of documents is proposed that is determined by the maximum correlation between explicit and implicit connectivity of the documents. A new science contextual citation index (SCCI) that is defined by a correlation maximum with a science citation index (SCI) is proposed and implemented. SCCI can more accurately measure scientific impact, find significant documents and evaluate new articles with zero SCI. The significant similar articles confirm each other and form clusters in the cyberspace. The research front exists as a set of such clusters. The proposed cyberspace implemented by WebVR and interactive

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований, гранты 16-07-00756, 16-29-09527, 18-07-00909 и 18-07-01111.

3D graphics can be considered as a dynamic learning environment that is convenient for discovering new significant articles, ideas and trends.

Keywords*: *cyberspace, web-based 3D graphic, WebVR, scientific papers, semantic similarity, science contextual citation index, implicit links, information integration, visualization.

I. Введение

Кибернетические миры и виртуальная реальность (ВР) часто используются в образовательных и исследовательских целях. VR – это, по сути, компьютерная реальность, также называемая киберпространством, виртуальной средой, симуляциями и искусственными мирами [1].

В этой статье мы описываем семантический подход к визуализации 3D-киберпространства научных работ и их исследований с использованием веб-3D-графики. Наиболее цитируемые и значимые документы в этом киберпространстве представлены сферами большого размера, а расстояние между документами – пропорционально их смысловому сходству. В этом киберпространстве фронт исследований может существовать как набор кластеров значимых документов.

Концепция исследовательского фронта была первоначально введена Прайсом [15] как совокупность статей, которые активно цитируют ученые. Для ученых и аналитиков необходимо понимание динамики исследовательского фронта, чтобы они могли идентифицировать новые тенденции и внезапные изменения в теле научных знаний. В этой статье мы разрабатываем концепцию исследования, предлагая учитывать как явные, так и неявные связи между документами. Более подробно наша концепция исследования фронтов описана в разделе 2.

В этой статье мы предлагаем трехмерное пространство научных работ, аналогичное созвездию Chaomei Chen [2; 3; 4], но имеющее улучшенные функции визуализации и смысловую точность. В последние годы был достигнут значительный прогресс как в разработке новых инструментов визуализации, таких, как WebVR [8], так и в анализе семантического подобия. Интерес к смысловому текстовому сходству постоянно растет. Например, Международный семинар по семантической оценке 2016 года увеличил число команд на 45% по сравнению с 2015 годом. На этой конференции методы Word2Vec [9; 11] и WMD [12] часто использовались для оценки семантического подобия текстов [22–24]. Метод Word2Vec оценивает семантическое сходство разных слов, в то время как метод WMD способен оценивать семантическое сходство разных фраз без общих слов. Метод Doc2Vec [10] позволяет встраивать слова и документы в общее семантическое векторное пространство.

Чтобы найти наиболее важные научные статьи, широко используется индекс научного цитирования (SCI). У SCI много недостатков, и он не подходит для изучения новых статей, которые представляют наибольший интерес для ученых. Чтобы преодолеть эту проблему, мы предлагаем новую меру, называемую индексом контекстного научного цитирования (SCCI), описанного в разделе 6. Также в этом документе предлагается новый подход для измерения семантического подобия документов, описанный в разделе 5.

Предлагаемое семантическое киберпространство научных работ, реализованное в WebVR и посредством интерактивной 3D-графики, можно рассматривать как динамичную среду обучения, которая удобна для обнаружения новых значимых статей, идей и тенденций.

Кроме того, следует отметить, что предлагаемое киберпространство может быть мощным инструментом интеграции информации, поскольку позволяет визуализировать документы разных языков в одном пространстве, если мера семантического сходства достаточно развита и поддерживает многоязычие.

II. Визуализация исследовательских направлений

Концепция исследовательского фронта была первоначально введена Прайсом [15] как совокупность статей, которые активно цитируют ученые.

Если мы определяем исследовательский фронт как текущее состояние специальности (т.е. линию исследований), то исследовательский фронт формирует его интеллектуальную базу. Специальность может быть концептуализирована как временное отображение $\Phi(t)$ исследуемого фронта $\Psi(t)$ в интеллектуальную базу $\Omega(t)$ [14].

В нашем подходе мы далее разрабатываем концепцию исследовательского фронта и предлагаем учитывать как неявные, так и формальные ссылки между статьями. Неявные ссылки вычисляются на основе меры семантического сходства, описанной в разделах 5 и 6. Неявная ссылка – это ссылка на автора или его идеи в тексте. Идея представляется как набор фраз с аналогичным значением. Фразы, представляющие идеи в текстах статей, образуют вместе систему неявных ссылок. Использование неявных ссылок вместе с формальными ссылками делает алгоритм более чувствительным и позволяет более детально исследовать фронт исследования.

III. Визуализация многомерных данных

В этой статье мы предлагаем трехмерное семантическое киберпространство научных работ, в котором документы представлены сферами, а расстояние между документами – пропорционально их смысловому сходству. Мы также предлагаем новую меру семантического сходства документов. Если мы вычислим семантическое сходство между всеми документами (N) в корпусе, то получим матрицу $N \times N$ взаимных расстояний. Для визуализации документов в трехмерном пространстве нам нужно преобразовать эту матрицу в размер $N \times 3$ и создать трехмерную визуальную карту. Это отображение матрицы взаимных расстояний в 3D неизбежно приводит к искажениям (неточностям) для размеров матриц более 4×4. Эта проблема правильной визуализации семантического подобия из-за большой размерности семантического векторного пространства может быть частично решена специальными методами уменьшения размерности и метода t-SNE.

В течение последних нескольких десятилетий было предложено множество методов визуализации многомерных данных. T-SNE – один из самых популярных алгоритмов сокращения размерности сегодня. Существует много модификаций, описанных в Интернете. T-SNE сохраняет большую часть локальной структуры многомерных данных, а также раскрывает глобальную структуру. Этот метод широко используется, поскольку после моделирования похожие объекты располагаются близко друг к другу, в то время как менее похожие – далеко друг от друга.

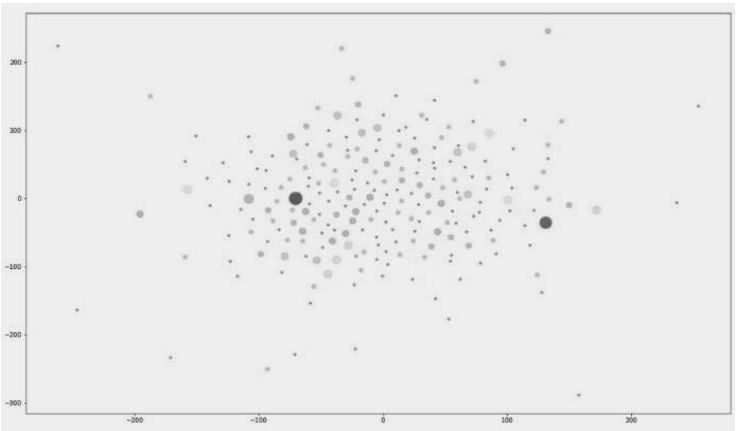

Рис. 1. 2D-карта визуализации корпуса текстов

В зависимости от размера визуализации (2D, 3D) используются разные методы. Если визуальная карта плоская, достаточно использовать стандартные инструменты визуализации Python, например библиотеку matplotlib.

В этой статье рассматривается корпус документов по компьютерной графике с заранее рассчитанной оценкой и мерой сходства. Корпус состоит из 230 статей. Каждая статья представлена в виде вектора смысловых мер сходства с другими статьями. Рейтинг каждой статьи равен числу цитат из других статей. Чем больше рейтинг (количество цитат) статьи, тем больше размер точки. Полученная двумерная визуальная карта статей представлена на рис. 1.

IV. WebVR визуализация 3D-карты

Мы используем технологию WebVR для создания трехмерной визуальной карты корпуса компьютерной графики на основе результатов алгоритма *t*-SNE. Технология WebVR позволяет поворачивать 3D-модель, просматривать ее под разными углами и тем самым создавать более полную картину структуры текстового корпуса.

Команда Mozilla VR разработала A-Frame в середине 2015 года. A-Frame – это структура WebVR, которая упрощает и ускоряет реализацию виртуальной реальности, позволяя вам кодировать HTML без необходимости знать мощный, но сложный WebGL [21]. A-Frame является открытым исходным кодом и в основном поддерживается Mozilla и сообществом WebVR. Это система компонентов компонента для Three.js, где разработчики могут создавать сцены 3D и WebVR с использованием HTML.

На рис. 2 представлена трехмерная визуальная карта корпуса из 230 российских документов по компьютерной графике. Эта карта является одной из фотографий сцены виртуальной реальности, которая была построена с использованием технологий WebVR и A-Frame.

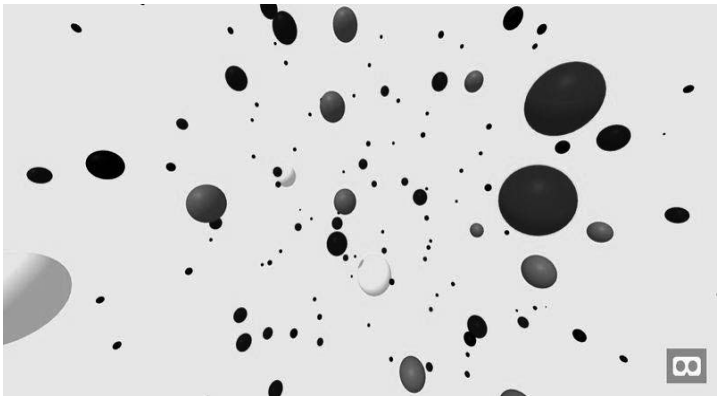

Рис. 2. 3D-визуальная карта корпуса
из 230 российских документов по компьютерной графике

V. Новая мера семантического сходства

В этой статье мы предлагаем новую меру семантического сходства, основанную на соотношении явных и неявных связей. Эта корреляция была впервые обсуждена в работе «Об основных типах связи между текстовыми документами» [6], где рассматривается вопрос о связности двух текстов естественного языка на основе их текстовых признаков (фрагментов). Выявлены два типа связности: явная (формальная) связь, когда тексты связаны библиографическими ссылками, и неявная связь, когда тексты связаны через общие текстовые фрагменты. Эксперимент проводился с использованием корпуса компьютерной графики из 120 связанных статей; каждая из них

цитируется или имеет библиографическую ссылку на другую статью в этом корпусе. На основе эксперимента показано, что оба типа связности коррелированы. Оптимальные параметры обработки текста обнаруживаются, когда корреляция максимальна и достигает около 55%. Результаты эксперимента показаны на рис. 3, где вертикальная ось представляет собой корреляцию, тогда как горизонтальная ось представляет количество букв в пороге длины текстового фрагмента. Из рисунка видно, что корреляция достигает максимума на пороге 7.

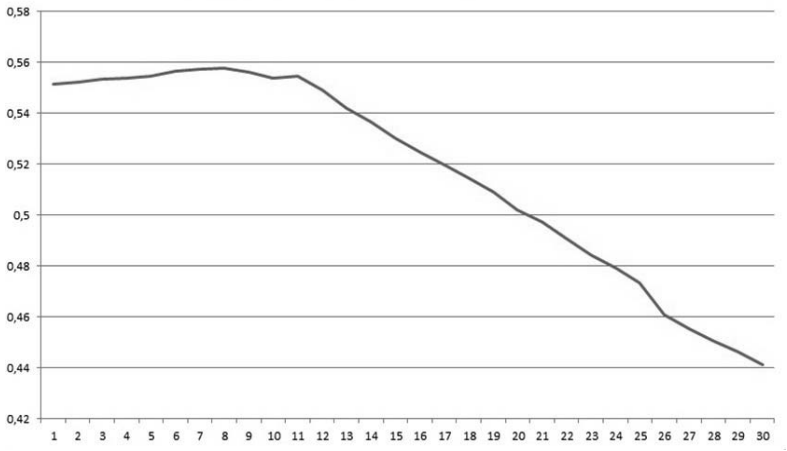

Рис. 3. Корреляция между явной и неявной связностью
в зависимости от порога длины текстового фрагмента

Преимущество такого подхода заключается в том, что он не требует ручной работы экспертов для оценки связности и сходства текстов. Вместо этого этот подход основан на более авторитетном мнении ученых, которое выражается в библиографических ссылках и в значении индекса научного цитирования.

VI. Индекс контекстного научного цитирования (SCCI)

В этой статье мы предлагаем Индекс контекстного научного цитирования (SCCI), который определяется количеством и качеством неявных ссылок на статью. Параметры алгоритма обнаружения неявных ссылок определяются максимальной корреляцией между явными и неявными ссылками. SCCI полезен для визуализации новых документов, так что более значимые и влиятельные документы могут быть представлены большими размерами. Обычно индекс научного цитирования (SCI) используется для обнаружения значимых научных документов, но SCI не может анализировать новые документы с нулевым SCI. Поэтому необходимость совершенствования SCI для наших целей самоочевидна. SCCI способен анализировать новые документы, потому что он более чувствителен, чем SCI. SCCI рассчитывается путем анализа текста статьи и сравнения его с текстами других статей с использованием меры семантического сходства. Алгоритм вычисления SCCI строится путем выбора оптимальных параметров обработки текста, которые обеспечивают максимальную корреляцию между SCCI и SCI. Таким образом, можно сказать, что SCCI является дальнейшим развитием SCI. Первая реализация SCCI уже выполнена и проверена с использованием корпуса компьютерной графики.

Таким образом, использование SCCI помогает нам решить задачу построения более реалистичных кибермиров, найти более важные статьи и более точно отразить семантическое сходство между ними.

VII. Выводы

В статье описывается семантический подход к визуализации 3D-киберпространства научных работ и их исследовательского фронта с использованием веб-3D-графики. Предложена модель киберпространства, в которой научные статьи представлены сферами разных размеров в соответствии с их значением и влиянием. Эта модель позволяет построить более реалистичное киберпространство научных работ для улучшения и облегчения научных исследований и разработки новых возможностей в динамичной учебной среде.

Благодарности.

Мы благодарны Российскому фонду фундаментальных исследований за финансовую поддержку наших проектов.

Литература

1. *Patel, H., Cardinali, R.* Virtual Reality Technology in Business // Management Decisio. – 1994. – Vol. 32. – Issue: 7. – Pp. 5–12.

2. *Chen, C.* Structuring and visualizing the WWW with Generalized Similarity Analysis // Proceedings of the Eighth ACM Conference on Hypertext (Hypertext'97). – Southampton. – 1997. – June. – UK. – Pp. 177–186.

3. *Chen, C.* Visualization of knowledge structures. Handbook of Software Engineering and Knowledge Engineering, 2002.

4. *Dodge, M., Kitchin, R.* Mapping Cyberspace. – Routledge, London, 2000.

5. *Grobelnik, M., Mladenić, D.* Visualization of news articles // Informatica Journal. – 2004. – Vol. 28. – No. 4. – Pp. 375–380.

6. *Charnine, M., Somin, N.* On the main types of connectivity between text documents // Systems and Means of Informatics. – 2017. – V. 27. – № 1.

7. *Neelakantam, S., Pant, T.* Learning Web-based Virtual Reality: Build and Deploy Web-based Virtual Reality Technology, 2017.

8. *Mikolov, Tomas, Corrado, Greg, Chen, Kai, Dean, Jeffrey*. Efficient Estimation of Word Representations in Vector Space // Proceedings of the International Conference on Learning Representations (ICLR 2013). – 2013. – Pp. 1–12.

9. *Lau, Jey Han, Baldwin, Timothy*. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, 1st Workshop on Representation Learning for NLP. – Berlin, Germany, 2016. – Pp. 78–86.

10. *Abdelwahab O., Elmaghraby A.* UofL at SemEval-2016 Task 4: Multi Domain word-2vec for Twitter Sentiment Classification. – SemEval-2016. – Pp. 169–175.

11. *Tian, Junfeng, Lan, Man.* ECNU at SemEval-2016 Task 1: Leveraging Word Embedding from Macro and Micro Views to Boost Performance for Semantic Textual Similarity. – SemEval-2016. – Pp. 618–624.

12. *Charnine, M., Klimenko*, *S.* Measuring of "Idea-based" Influence of Scientific Papers // Proceedings of the 2015 International Conference on Information Science and Security (ICISS, Seoul, December 14–16, 2015). – Seoul, South Korea, 2015. – Pp. 160–164.

13. *Charnine, M., Klimenko*, *S.* Semantic cyberspace of scientific papers // Proceedings of the 2017 International Conference on Cyberworlds (20–22 September, 2017). – Chester, United Kingdom, 2017. – Pp. 146–149.

14. *Synnestvedt, M.B., Chen C., Holmes J.H.* CiteSpace II: Visualization and Knowledge Discovery in Bibliographic Databases // AMIA 2005 Symposium Proceedings. – Pp. 724-728. Drexel University Libraries www.library.drexel.edu

15. *Price, DD.* Networks of scientific papers // Science. – 1965. – Jul 30. – Pp. 149:510-5.

16. *Morris, S.A., Yen, G., Wu, Z., & Asnake, B.* Time line visualization of research fronts // Journal of the American Society for Information Science and Technology. – 2003. – No. 54 (5). – Pp. 413–422.

17. *Erten, C., Harding, P.J., Kobourov, S.G., Wampler, K., & Yee, G.* Exploring the computing literature using temporal graph visualization (Tech. Rep. TR0304), University of Arizona, 2003.

18. *Boyack, K.W., Wylie, B.N., & Davidson, G.S.* Domain visualization using VXinsight for science and technology management // Journal of the American Society for Information Science and Technology. – 2002. – No 53 (9). – Pp. 764–774.

19. *Chen, C.* CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature // Journal Of The American Society For Information Science And Technology. – 2006. – No 57 (3). – Pp. 359–377.

20. *Chen, C., Ibekwe-SanJuan, F., Hou, J.* The Structure and Dynamics of Cocitation Clusters: A Multiple-Perspective Cocitation Analysis // Journal Of The American Society For Information Science And Technology. – 2010. – No 61 (7). – Pp. 1386–1409.

21. *Neelakantam, S., Pant, T.* Introduction to A-Frame // Learning Web-based Virtual Reality. Apress, Berkeley, CA., 2017.

22. *Zolotarev, O.V., Charnine, M.M., Matskevich, A.G., Kuznetsov, K.I.* Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet // Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015). – WORLDCOMP-15. – Las Vegas, Nevada, USA. – 2015. – July 27–30. – Vol. I. – Pp. 295–299.

23. *Galina1, I.V., Charnine, M.M., Somin, N.V., Nikolaev, V.G., Morozova, Yu.I., Zolotarev O.V.* Method for Generating Subject Area Associative Portraits: different Examples // Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015). – WORLDCOMP-15. – Las Vegas, Nevada, USA. – 2015. – July 27–30. – Vol. I. – Pp. 288–294.

24. *Zolotarev O., Charnine M., Matskevich A.* A Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts // Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014). – WORLDCOMP'14. – Las Vegas, Nevada, USA, CSREA Press. – 2014. – July 21–24. – Vol. I. – Pp. 82–87.

References

1. *Patel, H., Cardinali, R.* Virtual Reality Technology in Business // Management Decisio. – 1994. – Vol. 32. – Issue: 7. – Pp. 5–12.

2. *Chen, C.* Structuring and visualizing the WWW with Generalized Similarity Analysis // Proceedings of the Eighth ACM Conference on Hypertext (Hypertext'97). – Southampton. – 1997. – June. – UK. – Pp. 177–186.

3. *Chen, C.* Visualization of knowledge structures. Handbook of Software Engineering and Knowledge Engineering, 2002.

4. *Dodge, M., Kitchin, R.* Mapping Cyberspace. – Routledge, London, 2000.

5. *Grobelnik, M., Mladenić, D.* Visualization of news articles // Informatica Journal. – 2004. – Vol. 28. – No. 4. – Pp. 375–380.

6. *Charnine, M., Somin, N.* On the main types of connectivity between text documents // Systems and Means of Informatics. – 2017. – V. 27. – № 1.

7. *Neelakantam, S., Pant, T.* Learning Web-based Virtual Reality: Build and Deploy Web-based Virtual Reality Technology, 2017.

8. *Mikolov, Tomas, Corrado, Greg, Chen, Kai, Dean, Jeffrey*. Efficient Estimation of Word Representations in Vector Space // Proceedings of the International Conference on Learning Representations (ICLR 2013). – 2013. – Pp. 1–12.

9. *Lau, Jey Han, Baldwin, Timothy*. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, 1st Workshop on Representation Learning for NLP. – Berlin, Germany, 2016. – Pp. 78–86.

10. *Abdelwahab O., Elmaghraby A.* UofL at SemEval-2016 Task 4: Multi Domain word2vec for Twitter Sentiment Classification. – SemEval-2016. – Pp. 169–175.

11. *Tian, Junfeng, Lan, Man.* ECNU at SemEval-2016 Task 1: Leveraging Word Embedding from Macro and Micro Views to Boost Performance for Semantic Textual Similarity. – SemEval-2016. – Pp. 618–624.

12. *Charnine, M., Klimenko*, *S.* Measuring of "Idea-based" Influence of Scientific Papers // Proceedings of the 2015 International Conference on Information Science and Security (ICISS, Seoul, December 14–16, 2015). – Seoul, South Korea, 2015. – Pp. 160–164.

13. *Charnine, M., Klimenko*, *S.* Semantic cyberspace of scientific papers // Proceedings of the 2017 International Conference on Cyberworlds (20–22 September, 2017). – Chester, United Kingdom, 2017. – Pp. 146–149.

14. *Synnestvedt, M.B., Chen C., Holmes J.H.* CiteSpace II: Visualization and Knowledge Discovery in Bibliographic Databases // AMIA 2005 Symposium Proceedings. – Pp. 724-728. Drexel University Libraries www.library.drexel.edu

15. *Price, DD.* Networks of scientific papers // Science. – 1965. – Jul 30. – Pp. 149:510-5.

16. *Morris, S.A., Yen, G., Wu, Z., & Asnake, B.* Time line visualization of research fronts // Journal of the American Society for Information Science and Technology. – 2003. – No. 54 (5). – Pp. 413–422.

17. *Erten, C., Harding, P.J., Kobourov, S.G., Wampler, K., & Yee, G.* Exploring the computing literature using temporal graph visualization (Tech. Rep. TR0304), University of Arizona, 2003.

18. *Boyack, K.W., Wylie, B.N., & Davidson, G.S.* Domain visualization using VXinsight for science and technology management // Journal of the American Society for Information Science and Technology. – 2002. – No 53 (9). – Pp. 764–774.

19. *Chen, C.* CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature // Journal Of The American Society For Information Science And Technology. – 2006. – No 57 (3). – Pp. 359–377.

20. *Chen, C., Ibekwe-SanJuan, F., Hou, J.* The Structure and Dynamics of Cocitation Clusters: A Multiple-Perspective Cocitation Analysis // Journal Of The American Society For Information Science And Technology. – 2010. – No 61 (7). – Pp. 1386–1409.

21. *Neelakantam, S., Pant, T.* Introduction to A-Frame // Learning Web-based Virtual Reality. Apress, Berkeley, CA., 2017.

22. *Zolotarev, O.V., Charnine, M.M., Matskevich, A.G., Kuznetsov, K.I.* Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet // Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015). – WORLDCOMP-15. – Las Vegas, Nevada, USA. – 2015. – July 27–30. – Vol. I. – Pp. 295–299.

23. *Galina1, I.V., Charnine, M.M., Somin, N.V., Nikolaev, V.G., Morozova, Yu.I., Zolotarev O.V.* Method for Generating Subject Area Associative Portraits: different Examples // Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015). – WORLDCOMP-15. – Las Vegas, Nevada, USA. – 2015. – July 27–30. – Vol. I. – Pp. 288–294.

24. *Zolotarev O., Charnine M., Matskevich A.* A Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts // Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014). – WORLDCOMP'14. – Las Vegas, Nevada, USA, CSREA Press. – 2014. – July 21–24. – Vol. I. – Pp. 82–87.