Программы для наукометрических и библиометрических исследований: краткий обзор и сравнительный анализ

© Н.А. Мазов
Институт нефтегазовой геологии и геофизики им. академика А.А. Трофимука СО РАН, Новосибирск
MazovNA@ipgg.sbras.ru

© В.Н. Гуреев
Государственный научный центр вирусологии и биотехнологии «Вектор», Новосибирская область, Кольцово
gureyev@vector.nsc.ru

Аннотация

В последние годы в информационной практике наблюдается возрастающий интерес, привлекаемый к наукометрическим и библиометрическим исследованиям. Отчасти это связано с тем, что накоплены колоссальные объемы библиографической информации различного вида, требующей качественно новых форм аналитико-синтетической обработки, а с другой стороны это связано с более открытым и публичным доступом к наукометрическим базам данных. Несмотря на то, что лидеры производства и исследований наукометрических баз данных – Thomson Reuters (БД Web of Knowledge) и Elsevier (БД Scopus) – предоставляют необходимый сервис для анализа публикаций, тем не менее он остается весьма скудным и ограниченным. В настоящей работе представлен краткий обзор и дан сравнительный анализ программного обеспечения, предназначенного для различных аспектов наукометрических и библиометрических исследований и картографирования науки.

Подбор программного обеспечения для успешного решения информетрических исследовательских задач в области наукометрии и библиометрии является очень важным вопросом. В настоящее время существуют мощные коммерческие наукометрические программы с закрытыми исходными кодами. В то же время существует большое количество свободно распространяемых программ, как с открытыми, так и с закрытыми кодами [1]. В статье приводятся небольшие обзоры свободно распространяемых программ, предназначенных для наукометрических и библиометрических исследований, а также для картографирования науки.

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

1. Обзор программ для библиометрических исследований и картографирования науки

Мы представим двенадцать программ, специально созданных для библиометрического анализа научных отраслей и построения карт науки. Ниже приводится список этих программ.

- IN-SPIRE (1999)
- VantagePoint (2004)
- HistCite (2004)
- BibExcel (2009)
- CiteSpace II (2004)
- Sci^2 Tool (2009)
- Leydesdorff's Software (2004)
- Publish or Perish
- VOSViewer (2010)
- InterDisciplinary Research
- Network Workbench Tool (2007)
- SciMAT (2011)

Основное их отличие от инструментов, предлагаемых в Web of Science и Scopus, заключается в значительно более узкой специализации именно на анализе результатов и их визуализации, в то время как для продуктов Thomson Reuters и Elsevier это периферийные опции, достаточно грубо обрабатывающие библиографические данные. В то же время стоит упомянуть специализированную разработку Elseveir – SciVal [2], запущенную в 2011 г. и работающую с данными Scopus. К сожалению, ввиду дороговизны продукта (порядка 3 млн. рублей в год для наших организаций) мы не имели возможности оценить этот сервис. Между тем можно отметить его направленность на узкий круг пользователей, а именно – научные организации. Таким образом, продукт, в отличие от других перечисленных выше программ, остается недоступным рядовым ученым или научным группам.

1.1 Коммерческие программы

IN-SPIRE™ Visual Document Analysis [3] – коммерческое программное обеспечение для исследования и визуализации текстовых данных и аналитический инструментарий для получения

различных временных трендов. Используется для исследования научно-технической литературы.

VantagePoint [4] – мощный инструментарий для глубокого анализа текста с целью обнаружения знаний в результатах поиска в библиографических и патентных базах данных, позволяет быстро проанализировать большие объемы информации, отобранные в результате поиска, и превратить отобранную информацию в знание. Также является коммерческим продуктом.

1.2 Свободно распространяемые программы

Ниже дано краткое описание программных продуктов, распространяемых свободно.

HistCite [5] – очень гибкий программный продукт, создание и развитие которого инициировал основатель Института научной информации США и изобретатель Science Citation Index® доктор Юджин Гарфилд. Настоящий программный продукт предназначен для анализа и визуализации результатов поиска в БД Web of Science (файл в формате plain text) по различным критериям: по ключевым словам, авторам и цитируемым авторам, журналам, странам и организациям, от которых авторы публикуют свои статьи, и др. Программа легка и прозрачна в установке и использовании.

BibExcel [6] – программа предназначена для анализа библиографических данных или любых данных текстовой природы, отформатированной соответствующим образом. Идея состоит в том, что необходимо подготовить файлы данных, которые могут быть импортированы в Excel или любую программу, работающую с табличными данными, для дальнейшей обработки. Эта программа включает множество инструментальных средств, часть которых видима в окне, а часть других скрыта в меню. В качестве исходного файла для библиометрического исследования может быть использован файл результатов поиска в БД Web of Science, сохраненный в формате comma delimited. Программа разработана шведским ученым Persson O.D.

CiteSpace [7] – свободно распространяемое программное обеспечение для визуализации и анализа тенденций и направлений в научной литературе. Приложение разработано в виде инструмента для визуализации прогрессивной области знаний и направлено на поиск критических точек в развитии области или сферы, особенно интеллектуальных критических точек и важнейших моментов. CiteSpace имеет набор функций, облегчающих понимание и способствующих интерпретации сетевой или исторической модели, включая определение быстро растущих предметных областей, обнаружение «горячих точек» цитирования в области публикаций, разделения сети на кластеры, автоматическое маркирование кластеров в терминах, взятых из цитирующих статей, геопространственные модели совместных проектов, уникальные области международных совместных проектов. CiteSpace поддерживает структурный и временной анализ различных сетей, возникающих из научных публикаций, включая сети совместных проектов, сетей авторского социтирования и социтирования документов. Первичным источником входящих данных для CiteSpace являются результаты поиска в БД Web of Science, сохраненные в файле. CiteSpace предоставляет также несколько простых интерфейсов для получения данных из других систем, например из PubMed. Приложение можно использовать для генерации слоя географической карты, основанного на местоположениях авторов, который в дальнейшем можно посмотреть в Google Earth.

Science of Science (Sci^2) Tool [8] является блочным набором инструментов, специально разработанным для изучения науки. Он поддерживает временной, геопространственный, тематический и сетевой анализ и визуализацию набора данных на различных уровнях (микро – индивидуальном, мезо – локальном и макро – глобальном). Пользователи могут:

- воспользоваться онлайновыми наборами данных по науке либо загрузить свои собственные данные;
- проводить различные типы анализа с наиболее эффективными доступными алгоритмами;
- использовать различные воплощения для интерактивного объяснения и понимания специфических наборов данных;
- распределять наборы данных и алгоритмы сквозь научные границы.

Sci^2 Tool построен на открытом программном обеспечении для простого объединения и использования наборов данных, алгоритмов, инструментов и вычислительных ресурсов.

Leydesdorff Software [9] представляет собой набор консольных программ открытого доступа для разбора, преобразования и анализа библиометрических данных, полученных из таких источников, как Scopus, Web of Science и Google Scholar. Можно проводить анализ на соавторство, сети совместных проектов между странами, организациями и городами, анализ на одинаковые ключевые слова, социтирование, библиографический анализ и пр. Хотя непосредственно в приложение и не включены инструменты визуализации, оно подготавливает данные для создания реляционной базы данных и визуализации в других программах.

Publish or Perish [10] – программное обеспечение для поиска и анализа научных цитирований. Оно использует Google Scholar для получения необработанных цитат, а затем анализирует их и выводит следующие статистические данные:

- общее число статей;
- общее число цитирований;
- среднее число цитирований на статью;
- среднее число цитирований на автора;
- среднее число статей на автора;
- среднее число цитирований в год;

- индекс Хирша с относящимися к нему параметрами;
- индекс Эгга;
- современный h-индекс;
- взвешенный по возрасту показатель цитирования;
- два варианта индивидуальных h-индексов.

Результаты доступны с экрана, их также можно скопировать в буфер обмена Windows (для использования в других приложениях) или сохранить в одном из нескольких выходных форматов (для последующего упоминания и дальнейшего анализа). В Publish or Perish включено детальное руководство пользователя с подсказками и дополнительной информацией о системе показателей цитирования.

VOSviewer [11] – программа открытого доступа, которую можно использовать для различных целей. VOSviewer можно использовать для создания карт, основанных на сети данных. Карты создаются с использованием технологии создания карт VOS и технологии выделения кластеров VOS. VOSviewer можно использовать для обозрения и исследования карт. Программа выводит карту различными способами, каждый из которых выделяет ее различные аспекты. Предложены такие функции, как увеличение, прокрутка и поиск, которые облегчают тщательное исследование карты.

Изначально VOSviewer задумывался для анализа библиометрических сетей. Программа, например, может использоваться для создания карт по публикациям, статьям или журналам, основанных на сети социтирования, или создавать карты ключевых слов, основанных на их одновременном появлении в сети.

InterDisciplinary Research («IDR») [12] – данный веб-сайт предлагает новый инструмент – слои на карте науки в качестве метода исследования степени междисциплинарности набора публикаций. Технология наложения слоев показывает распространение публикаций на глобальной карте науки, т. е. структуре науки, какой она получается на основе анализа перекрестного цитирования между дисциплинами. Междисциплинарное исследование часто считается существенным для научного и технологического развития. Однако междисциплинарность – это неоднозначное и многоаспектное понятие: существует несколько допустимых способов ее выделения, и нет единого мнения по поводу того, какое из определений наиболее приемлемо.

Анализ можно провести на различных единицах объединения: например, для университета или корпорации, для темы исследования или для исследовательской программы или финансирующей организации. Располагая публикации на карте науки, можно понять разнообразие вовлеченных дисциплин. Научное приписывание публикаций к дисциплинам проблематично и противоречиво, карты со слоями надежны только тогда, когда на них много чисел.

Карты позволяют интуитивно понять различные аспекты дисциплинарного разнообразия. Во-первых, количество включенных дисциплин; во-вторых, соотношение дисциплин, т. е. распределены ли публикации равномерно или с преобладанием какой-либо дисциплины; в-третьих, и это важно, включено когнитивное расстояние между вовлеченными во взаимодействие дисциплинами – покрывает ли рассматриваемое исследование далекие или родственные области науки. Этот аспект несоответствия – ключевое преимущество карт: они проводят границу между междисциплинарностью малого радиуса (например, химия и физика) и большого радиуса (например, общественные науки и биология).

Network Workbench (NWB) [13] – поддерживает сеть научных исследований, преодолевая научные границы. Пользователи NWB имеют онлайновый доступ к большинству сетевых наборов данных или могут загружать собственные сети. Они имеют возможность представить сетевой анализ посредством наиболее эффективных доступных алгоритмов. Кроме того, у них имеется возможность создавать, запускать и проверять достоверность сетевых моделей для улучшения их представления о структуре и динамике специфических сетей. NWB предоставляет продвинутый инструментарий для визуализации, позволяющий интерактивно исследовать и понимать специфические сети, а также их взаимодействие с другими типами сетей.

SciMAT [14] – новая свободно распространяемая программа, предназначенная для проведения картографического анализа науки за многолетний период в хронологическом разрезе. В ней предусмотрены различные модули, позволяющие исследователю вести работу на различных этапах процесса составления карт науки. Модули программы помогают осуществлять каждый этап процесса картографирования науки. Выдающимися качествами программы являются наличие мощного модуля для предварительного приведения исходных библиографических данных к единому формату, возможность использования библиометрических индикаторов для анализа влияния каждого рассматриваемого элемента, а также большие возможности для конфигурирования анализа.

Несмотря на то, что программа изначально разрабатывалась для проведения концептуального картографического анализа науки, в дальнейшем она была расширена, чтобы стало возможным проводить любой тип картографического анализа науки (включая интеллектуальный и социальный). Это позволило определять имеющиеся подструктуры (главным образом, группы авторов, слов и ссылок) в области исследований средствами библиометрического анализа (библиографическая связка, журнальная библиографическая связка, авторская библиографическая связка, соавторы, социтирование, авторское социтирование, журнальное социтирование или анализ совместного появления слов) для каждого изучаемого периода.

2. Сравнительный анализ программ

Для проведения сравнительных исследований была проанализирована часть программ, кратко описанных ранее. Во внимание принимались следующие аспекты:

- методы предварительной обработки;
- доступные библиометрические сети;
- используемые методы нормирования;
- виды анализа;
- дополнительные параметры.

2.1 Методы предварительной обработки

Проведение предварительной обработки данных является важным свойством программ, предназначенных для библиометрических исследований и картографирования науки. Наиболее важными модулями предварительной обработки являются:

- устранение повторных записей;
- квантование времени;
- уплотнение данных.

Модуль определения повторных записей важен при анализе совместно появляющихся слов или соавторов. С его помощью пользователь может объединить два и более элементов, представляющих одно и то же понятие или одного и того же автора. Данный модуль не только объединяет два элемента, но и отбирает и суммирует значение параметра, например, число цитирований изначальных записей.

Опция квантования времени необходима в том случае, когда пользователю нужно проанализировать эволюцию изучаемой отрасли.

Модуль уплотнения данных необходим, когда пользователю необходимо отфильтровать данные для анализа наиболее значимой информации.

Только программы CiteSpace, VantagePoint, NWB, Sci^2 и SciMAT имеют наличие этих модулей предварительной обработки. Программы Leydesdorff's Software и VOSViewer не имеют ни одного из этих модулей, что представляется серьезным недостатком.

2.2 Библиометрические сети

При выборе программ для библиометрического анализа и картографирования науки необходимо знать, способны ли программы устанавливать различные связи между элементами анализа, т. е. способны ли они извлекать различные библиометрические сети.

Несмотря на то, что нет ни одной программы, в которой можно было бы построить все различные виды библиометрических сетей, программы BibExcel, CiteSpace, Leydesdorff's Software, Sci^2 Tool, VantagePoint и SciMAT способны построить большинство из них. Напротив, в VOSViewer невозможно построить ни одну сеть, она направлена лишь на графическое представление библиографических карт.

Некоторые программы позволяют создавать необычные сети. Например, сети по совместным появлениям грантов (CiteSpace), сети по совместным главным исследователям (Sci^2 Tool), а особые матрицы можно создавать в BibExcel и VantagePoint, используя набор определенных полей документов. Кроме того, некоторые программы (BibExcel и VantagePoint) позволяют получать разнородные сети посредством различных полей в строках и столбцах, например, матрицы, где показано количество авторов в расчете на годы. А в программах NWB Tool, Sci^2 Tool и SciMAT можно получать сети, используя прямую связь.

2.3 Показатели нормирования

После построения библиометрических сетей следует провести процесс нормирования, используя различные меры подобия. Часть из рассматриваемых программ (BibExcel, VantagePoint, Leydesdorff's Software и CiteSpace) используют в качестве меры подобия косинус Солтона. Другие программы (NWB Tool, Sci^2 Tool и SciMAT), позволяют пользователям задавать собственные показатели.

2.4 Виды анализа и прочие аспекты

Применяются различные методы анализа, доступные в каждой программе. Только в CiteSpace, Sci^2 Tool, VantagePoint и SciMAT используется наиболее широкий спектр методов анализа, в то время как Leydesdorff's Software не проводит ни один из видов. В CiteSpace и Sci^2 Tool реализованы возможности геокодирования. Так, например, CiteSpace использует геокоды от Google для доступных данных по организации. А в Sci^2 Tool используется внутренний геокодировщик, применяемый для любых полей, содержащих географические данные, таких как адрес организации, место проведения конференции и др.

2.5 Дополнительные параметры

В этом параграфе мы сравним программы по другим параметрам, таким как документация и справка, ценовая доступность, доступность исходного текста программы и возможность установки программы на различных операционных системах, а также возможности расширения программы.

Только две из рассматриваемых нами программ являются коммерческими, как уже было отмечено выше: IN-SPIRE и VantagePoint. Все остальные программные продукты распространяются бесплатно.

NWB Tool, Sci^2 Tool и SciMAT снабжены подробным руководством пользователя, где доступно объясняется работа с программами. Кроме того, в руководстве пользователя объясняются важные моменты по картографированию науки. Практически для всех программ реализовано хорошее руководство пользователя и онлайн-поддержка.

Открытые исходные тексты программ доступны только для NWB Tool, Sci^2 Tool и SciMAT. Программы CiteSpace, NWB Tool, Sci^2 Tool, VOSViewer и SciMAT разработаны на языке

программирования Java, так что они могут использоваться в любой операционной системе (Windows, MacOS, Linux и др.). С другой стороны, BibExcel, CoPalRed, IN-SPIRE, Leydesdorff's Software и VantagePoint доступны только в среде ОС Windows.

Наконец, если рассматривать программы с точки зрения возможности их расширения, то NWB Tool, Sci^2 Tool и SciMAT могут расширяться с использованием платформы Java. VantagePoint может быть расширена с использованием скриптов VisualBasic.

Для завершения сравнительного исследования программного обеспечения мы провели библиометрический анализ науки с определенной единицей анализа. В качестве исходного массива были использованы данные, полученные из БД Web of Science по запросу «геоэлектрика». Подробные результаты анализа будут представлены в ближайших публикациях авторов.

Заключение

В заключение следует отметить, что представленный список программного обеспечения является практически исчерпывающим списком программ для библиографического анализа и картографирования науки, которые используются в настоящее время (не рассматривались программы, не имеющие англоязычных версий). Эти программы имеют различные характеристики; например, некоторые из них направлены лишь на графическое представление, тогда как другие обладают различными модулями предварительной обработки. Ни одну из рассмотренных программ нельзя признать лучшей. Исключение, пожалуй, составляет лишь программа SciMAT. Следовательно, мы считаем, что исчерпывающий библиометрический и картографический анализ определенной области науки должен проводиться с использованием нескольких из этих программ, что позволит собрать все важные знания с различных углов зрения. Такая кооперация программ дает положительный эффект, позволяющий получать знание, скрытое в данных. Рассмотренный нами набор программ может быть рекомендован специалистам в областях, связанных с библиометрическими исследованиями и исследованиями в области построения карт науки.

Литература

[1] Мазов Н.А. Свободно распространяемые программы для наукометрических и библиометрических исследований // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 19-я междунар. конф. "Крым 2012" (2–10 июня 2012 г., г. Судак): Труды конф. – М.: Изд-во ГПНТБ России, 2012. – С. 1–6.

[2] Vardell, E., Feddern-Bekcan, T., Moore, M. SciVal experts: A collaborative tool // Medical Reference Services Quarterly. – V. 30(3). – P. 283–294.

[3] IN-SPIRE™ [Электронный ресурс]. – Режим доступа: http://in-spire.pnnl.gov/ (Дата обращения: 13.05.2013).

[4] VantagePoint – Text Mining software for Technology Management – Search Technology, Inc. [Электронный ресурс]. – Режим доступа: http://www.thevantagepoint.com/ (Дата обращения: 13.05.2013).

[5] Thomson Reuters – HistCite – Science [Электронный ресурс]. – Режим доступа: http://www.histcite.com/ (Дата обращения: 13.05.2013).

[6] BibExcel [Электронный ресурс]. – Режим доступа: http://www8.umu.se/inforsk/Bibexcel/ (Дата обращения: 13.05.2013).

[7] CiteSpace: visualizing patterns and trends in scientific literature [Электронный ресурс]. – Режим доступа: http://cluster.cis.drexel.edu/~cchen/citespace/ (Дата обращения: 13.05.2013).

[8] [Электронный ресурс]. – Режим доступа: https://sci2.cns.iu.edu/ (Дата обращения: 13.05.2013).

[9] Sci^2 Tool: A Tool for Science of Science Research and Practice [Электронный ресурс]. – Режим доступа: http://www.leydesdorff.net/ (Дата обращения: 13.05.2013).

[10] Publish or Perish – Anne-Wil Harzing [Электронный ресурс]. – Режим доступа: http://www.harzing.com/pop.htm (Дата обращения: 13.05.2013).

[11] VOSviewer [Электронный ресурс]. – Режим доступа: http://www.vosviewer.com/ (Дата обращения: 13.05.2013).

[12] IDR – Interdisciplinary Research – Measuring and Mapping Interdisciplinary Research [Электронный ресурс]. – Режим доступа: http://idr.gatech.edu/ (Дата обращения: 13.05.2013).

[13] Network Workbench [Электронный ресурс]. – Режим доступа: http://nwb.cns.iu.edu/ (Дата обращения: 13.05.2013).

[14] SciMAT – Science Mapping Analysis Tool [Электронный ресурс]. – Режим доступа: http://sci2s.ugr.es/scimat (Дата обращения: 13.05.2013).

Software for scientometric and bibliometric research: a brief overview and comparative analysis

Nikolai A. Mazov, Vadim N. Gureyev

In the last few years we can observe growing interest to scientometric and bibliometric studies in the field of information science. To some extent it is caused by large

volumes of bibliographic information of different kinds. This information needs conceptually new methods of analytic-synthetic treatment. On the other hand, it is connected with access to scientometric databases that became more available and public. Whereas leading companies in producing of scientometric databases – Thomson Reuters (Web of Knowledge) and Elsevier (Scopus) – provide necessary services for publication analysis, these services are limited and rather poor. In this article we made a brief overview, as well as presented comparative analysis of modern software designed for studying of different scientometric and bibliometric aspects and for science mapping analysis.