OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822: 004.912

ПОСТРОЕНИЕ СЕТЕЙ СОАВТОРСТВА ПО ДАННЫМ СЕРВИСА GOOGLE SCHOLAR CITATIONS

Ландэ Д.В., Балагура И.В., Андрущенко В.Б.

Институт проблем регистрации информации НАН Украины,
г. Киев, Украина

dwlande@gmail.com

balaguraira@mail.ru

valentyna.andrushchenko@gmail.com

В работе приводится алгоритм построения сети соавторства ученых, регулируемой их научными интересами. Сеть соавторства формируется на основе зондирования сервиса GoogleScholarCitations.Показано, что дескрипторы, определяющие тематическую направленность, влияют на размер формируемой сети, а также на динамику ее роста. Показано, чтокластерыв сетях соавторства могут рассматриваться как основа для выявления научных школ.

Ключевые слова:vсеть соавторства; наукометрия;Google Scholarv Citations; предметная область; граф связей; зондирование сетей.

Введение

Развитие Интернета, поисковых систем и социальных сетей повлияло на многие сферы деятельности человека, в том числе и на развитие науки. Трансформация реферативных баз данных, возникновение различных научных платформ для распространения, обмена научной информацией спровоцировало новые подходы ведения научных исследований, формирования научных групп и поиска возможного сотрудничества. Как следствие развития сервисов научной информации, появились новые возможности оценки научной информации и изучение закономерностей научного взаимодействия [Ortega, 2015]. Основным инструментом изучения закономерностей научного сотрудничества являются сети соавторов, с помощью которых можно получить не только наукометрические оценки, но и определять экспертов для решения сложных заданий. Одним из крупных сервисов научной информации является GoogleScholar, который позволяет создавать ученым профили, содержащие соответствующую библиографическую информацию, а также осуществлять поиск публикаций со всего мира. Изучению сетей соавторов так же как и сервисаGoogle Scholar Citation (http://scholar.google.com/citations) посвящено большое количество работ, что подтверждает актуальность проводимых исследований [Liu, 2015]. Среди них методы построения сетей соавторов, определения значимых узлов, структуры сети, исследование цитирования в GoogleScholar, а так же соответствующих корпусовдр. [Brezina,2012].

Предлагается методика построения сетей соавторства – моделей сотрудничества ученых на основе зондирования наукометрических сетей. Как такая сеть в работе рассматривается сеть понятий, соответствующих тегам сервиса GoogleScholarCitations.

Целью работы является описание теоретических принципов и методологии автоматизированного формирования сетей соавторства, в частности, областях ComplexNetworksи TextMining путем зондирования большой информационной сети. Для достижения этой цели применяется специальный алгоритм сканирования ресурсов сервиса Google Scholar Citations с целью получения репрезентативного набора соавторов как основы (узлов) будущей сети. Под зондированием сетей будем понимать выборку небольшого объема важнейшего содержания из больших сетей, которые по технологическим причинам не подлежат полному сканированию[Lande, 2015].

Очевидно, сеть соавторства может иметь достаточно большие размеры, если ее не ограничивать определенной тематикой, например, задаваемой тематикой первого автора, начиная с которого идет процесс формирования этой сети.

Данный эффект значительно усложняет восприятие сформированной сети и приводит к такому эффекту, как «дрейф тематики». Также имеет место одинаковое написание фамилий и инициалов различных ученых. Для преодоления этих эффектов применяется тематическая фильтрация, т.е. используются дескрипторы, приписываемые авторам наукометрической сети, определяющие их тематическую направленность. Соответствие этим дескрипторам в конечном итоге иопределяют размер формируемых сетей соавторства, а такжединамику ее роста. Кроме того, распознавание кластеров в таких сетях может рассматриваться как основа для выявления научных школ, экспертных групп и т.п.[Landeetc, 2013].

1. Описание модели

При построении сетей соавторства целесообразно применять модели, уже апробированные на пиринговых сетях (peertopeer, P2P – равный с равным), основанных на равноправии участников. Пиринговые сети состоят из узлов, каждый из которых взаимодействует лишь с некоторым подмножеством других узлов, что вполне соответствует сети соавторов.

Для поиска необходимых данных в таких сетях применяется несколько моделей. В модели "широкого первичного поиска" (Breadth First Search, BFS) запрос из некоторого стартового узла адресуется ко всем соседям (ближайшим по некоторым критериям)[Kalogeraki, 2002]. Когда некоторый другой узел получает запрос, выполняется поиск в его локальном индексе и в случае успеха возвращает результат. В противном случае запрос передается по сети далее. В нашем случае (зондирование сети соавторов) запрос передается далее во всех случаях, если не выполняются некоторые условия-ограничения.

Зондирование опорной модельной сети осуществляется по такому алгоритму [Lande, 2015]:

1. Выбирается определенное количество узлов опорной (зондируемой) сети, определяемых как базовые для новой сети, соответствующей результатам зондирования (в простейшем случае выбирается один узел).
2. Для каждого из рассматриваемых узлов опорной сети определяются смежные с ним узлы (соавторы), которые добавляются к создаваемой сети как результаты зондирования. Формируются ребра-связи к этим узлам из исходного узла.
3. От текущего узла опорной сети осуществляется переход к случайно выбранному соседнему узлу формируемой сети (соавтору).
4. Если имеет место "зацикливание" (выбирается узел, к которому уже был осуществлен переход по этому алгоритму) или несоответствие узла некоторому условию-ограничению, происходит переход к другому случайно выбранному узлу формируемой сети. Если таких узлов не осталось, считается, что сеть, соответствующая результатам зондирования, построена

В качестве условий-ограничений на практике проверяется вхождение в теги соавтора некоторого множества допустимых. В рамках модели просто каждому из узлов с некоторой вероятностью приписывается возможность порождения от него последующих связей.

При моделировании приведенный алгоритм применялся для двух самых распространенных модельных сетей Erdős-Rényi (ER) и Barabási-Albert (BA). Известно, что модель ER [Erdős, 1960] – это случайная сеть, которая строится следующим образом: множество из N изначально не соединенных узлов попарно объединяют с вероятностью p. В результате создается сеть приблизительно с $pN(N-1)/2$ случайно выбранными связями.

Модель BA [Barabási, 1997] –сеть со степенным распределением степеней узлов (так называемых, безмасштабных сетей). Эта модель учитывает принцип преимущественного присоединения, который заключается в том, что чем больше связей имеет узел, тем более вероятно для него создание новых связей со вновь образуемыми узлами.

Следует отметить, что безмаштабными являются наиболее популярные реальные сети, такие как веб-пространство с гиперссылками, социальные сети, сети слов в литературных произведениях, сети протеинов, и т.п. [Newman, 2003]Авторами было показано, что сети соавторства тоже обладают свойством безмасштабности. Исходя из информации о том, что все известные большие сети цитирования, соавторства и т.п. обладают свойством безмасштабности, т.е. в чем-то близки по структуре сети Barabási-Albert, изучались модели, одна из которых базировалась на алгоритме BA. От этой модели принципиально отличаются случайные сети Erdős-Rényi, которые также изучались для сравнения. Сравнение показывает, что связанные области (ветки), соответствующие отдельным понятиям в случае модели ER достаточно длинные, а узлов, по которым следует маршрут зондирования больше, чем в более интересном случае модели BA (рис. 1). В данном случае нам важны именно качественные результаты, вид связанных цепочек, которыми моделируются ветки понятий. Следует отметить, что реальным сетям присущий еще и феномен "клуба богатых" (Rich Clube), который обусловливает более плотную связанность наибольших узлов-соавторов.

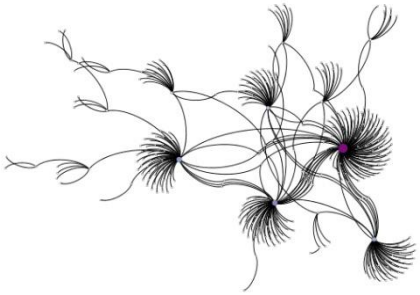


Рисунок 1 -- Пример сети, построенной зондированием сети типа Barabási-Albert

Именно на основании результатов качественного моделирования был сделан вывод о возможности формирования небольших связанных веток соавторов, соответствующих понятиям, интересующим пользователей сервиса GoogleScholarCitations.

2. Зондирование сети Google Scholar Citations

Приведенный алгоритм адаптировался к реальной сети соавторов сервиса Google Scholar Citations следующим образом (рис.2):

1. Выбирается первый автор, с которого начинается зондирование.
2. Экспертным путем определяется небольшой перечень базовых тегов-дескрипторов, соответствующих наиболее важным понятиям.
3. Открывается страница веб-сервиса, соответствующиевыбранному автору.
5. К создаваемой сети добавляются все соавторы, содержащиеся на странице выбранного автора. Формируются ребра-связи к этим узлам (соавторам) из исходного узла (автора).
4. Из списка узлов формируемой сети случайным образом тот, на страницу которого планируется перейти для дальнейшего анализа. Этот узел также должен удовлетворять тематике выбранной предметной области (его теги входят в состав дескрипторов, определенных на шаге 2) и не входит в состав тех узлов, к страницам которых уже был осуществлен переход.
5. Если такой узел-автор выбран, то происходит переход к пункту 3.
6. Если такого автора не существует, то считается, что сеть зондирования построена.

В соответствии с приведенным алгоритмом процесс зондирования сети, начиная с определенного узла, прекращается при «зацикливании», т.е. когда в соответствии с

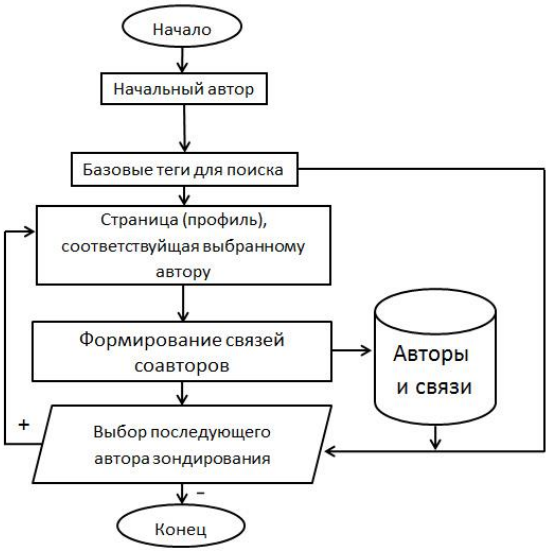


Рисунок 2 -- Алгоритм зондирования сети Google Scholar Citations

алгоритмом происходил переход к уже пройденному узлу, а также при отклонении оставшихся соседних узлов от основной тематики (определяется учетом лексического состава тегов). При этом само «зацикливание» является признаком перехода к следующему базовому автору или завершению процесса зондирования.

3. Примеры построения сетей соавторства

Построения в соответствии с приведенным алгоритмом построена сеть соавторов при достаточно широком списке дескрипторов-ограничений (computer, networks, language, information, complex, text) и ограничении на количество сканируемых узлов в 1000. С помощью программного средства получена визуализация данной сети соавторов (рис. 3).

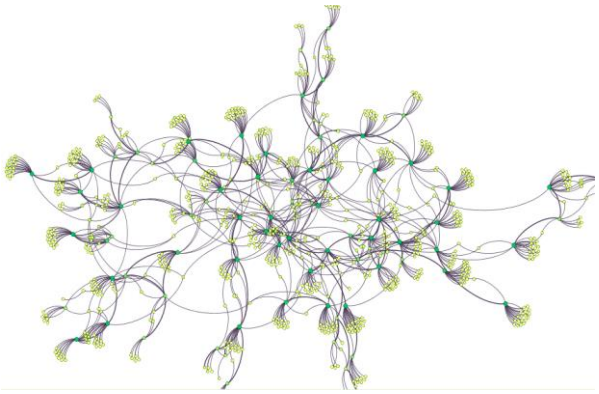


Рисунок 3 -- Фрагмент сети соавторов, построенная с учетом указанных широкой тематики дескрипторов

Как видно по рис. 4. динамика роста сети имеет четкий линейный тренд из-за чересчур широкой предметной области.

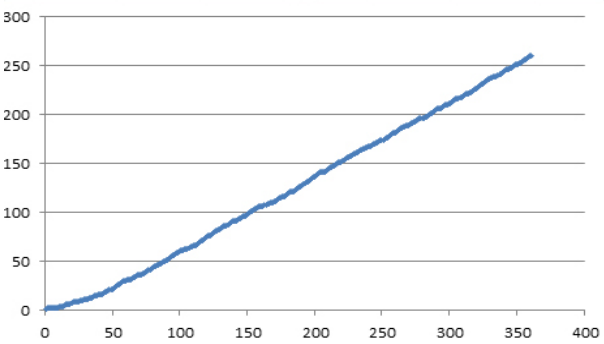


Рисунок 4 -- Динамика роста сети соавторов, построенная с учетом указанных дескрипторов (горизонтальная ось – шаги выбора узлов (авторов), вертикальная – количество узлов)

Для построения сетей соавторовпо тематике TextMining экспертным путем были определены базовые теги на английском языке: language processing, text mining, information retrieval, complex networks. На рис. 5 приведен пример сети соавторов, построенной в соответствии с приведенным алгоритмом по указанным базовым дескрипторам-тегам (первый узел зондирования соответствовалодному из авторов данной статьи).

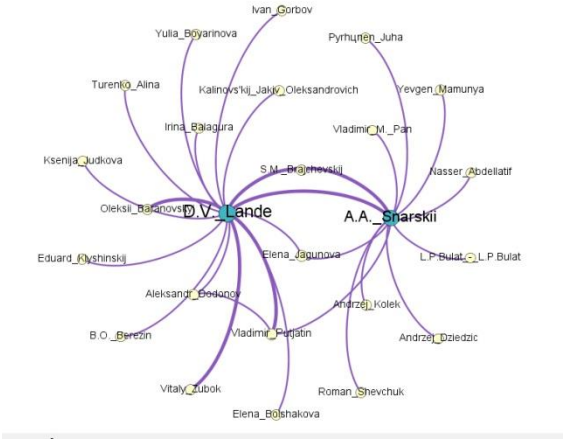


Рисунок5 -- Небольшая сеть соавторов, построенная с учетом указанных дескрипторов

Динамика роста данной сети при зондировании сервиса GoogleScholarCitations имеет вид, представленный на рис. 6.

Применение методов кластерного анализа позволяют выявлять наиболее тесно связанные между собой группы ученых-соавторов, научных школ, экспертных групп. В данном случае под научной школой будем понимать неформальный творческий коллектив исследователей разных поколений, объединенных общей программой и стилем исследовательской работы, которые действуют под руководством признанного лидера.На рис. 7 показано визуальное представление процесса выявления кластеров путем пошагового удаления наименее весомых ребер из сети соавторства, построенной по тематике ComplexNetworks.

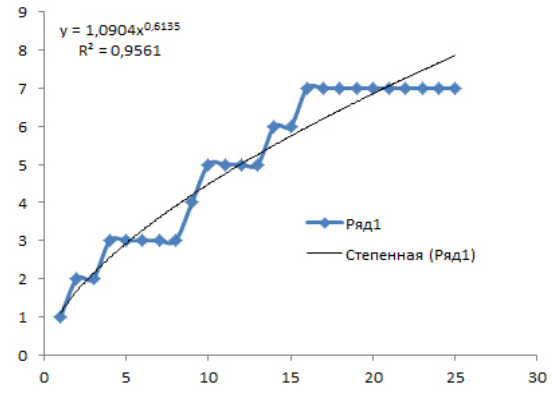
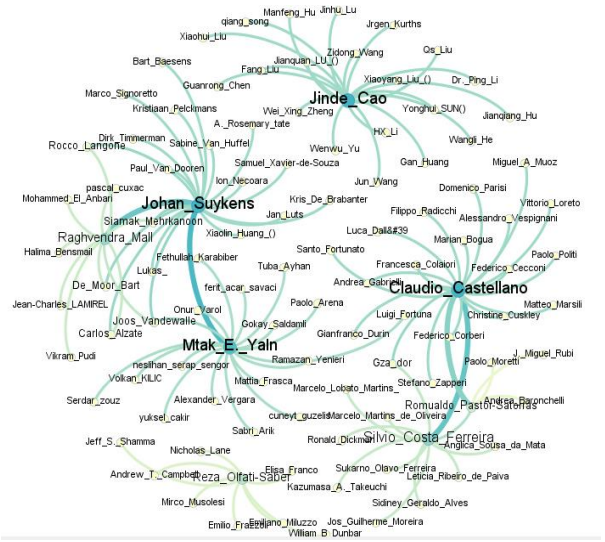


Рисунок6 -- Динамика роста сети соавторов, построенная с учетом указанных дескрипторов

В данном случае приведенный выше алгоритм применялся к группе из 10-и ученых, имеющих в соответствии с данными GoogleScholarCitations наибольшую цитируемость.



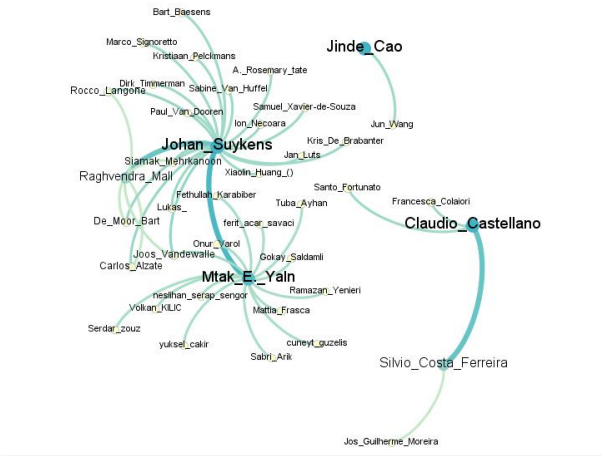


Рисунок 7 -- Кластеризация путем пошагового удаления ребер средствами программы Gephi

4. Методика построения модели предметной области

Дальнейший детальный наукометрический анализ предметной области позволяет выделить наиболее активных в научном сотрудничестве

ученых и научные группы для решения заданных вопросов.

Авторами была предложена и опробована методика наукометрического анализа, включающая методы анализа сложных сетей, методы фильтрации текста, визуализацию данных с помощью программных средств Gephi. На первом этапе определяется область и научные направления, по которым будет проведен анализ, загружается из сети (базы данных) и фильтруется файл с реферативной информацией. Результат первого этапа – отфильтрованные по определенным дескрипторам данные об авторах и связях между ними, т.е. матрица, соответствующая сети.

Второй этап – создание сети соавторов выбранной предметной области, а также основных характеристик сети с помощью программных средств Gephi, а также расчет дополнительных параметров спомощьюсобственных программных средств. В результатевторогоэтапаопределяются основные качества сотрудничества ученых, научные группыисамые коммуникативныеученые по определенному научному направлению.

Третий этап посвящен отбору полнотекстовых публикаций самых коммуникативных ученыхи создание текстового корпуса для выявления основных терминов (слов и словосочетаний) по научным направлениям.

На четвертом этапевыполняется визуализация сетей терминов ученых и области в целом, расчет основных параметров. Проводится обобщение результатов, описание основных характеристик, тенденций в области.

Результаты исследований дают возможность научно обосновать, автоматизировать и ускорить процедуру подбора компетентных экспертов для решения различных вопросов так и производства новой продукции.

Заключение

Предложен и реализован подход к формированию сетей соавторства в рамках предметной области, ограничительными элементами которого составляют некоторые маркеры знаний (теги), заранее заданные учеными – участниками проекта GoogleScholarCitations.

Следует отметить принципиальное отличие предложенной модели автоматического формирования сетей соавторства от существующих, базирующихся на непосредственном участии экспертов при выборе конкретных узлов и связей. В данном случае исследователь для построения сетииспользует лишь крупицы знаний, представленных в виде набора базовых тегов. В дальнейшем программа использует знания, заложенные соавторами, теги отмеченные как главные для них. Т.е. экспертная середа в этом случае существенно расширяется.

Модель применялась для отраслей науки ComplexNetworks и TextMining в рамках сервиса GoogleScholarCitations, но предложенный подход можно использовать и для других научных областей, или для других наукометрических массивов.

Результаты моделирования с помощью предложенной в разделе 3 методики также могут использоваться для создания модели предметной области, поиска групп экспертов. Выполнение подобных исследований могут способствовать усовершенствованию инструментария научных сервисов за счет внедрения технологий наукометрического анализа, улучшению эффективности аналитической деятельности.

Библиографический список

[Barabási, 1997] Barabási A., Albert R. Emergence of scalingin random networks // Science.– 1997. – 286. – P. 509-512.

[Brezina, 2012]Brezina V.Use of Google Scholar in corpus-driven EAP research// Journal of English for Academic Purposes.– 2012. –11. – P. 319-331

[Erdős, 1960]Erdős P., Rényi A. The Evolution of RandomGraphs // Magyar Tud. Akad. Mat. Kutató Int. Közl., 1960. – 5. –P. 17–61.

[**Kalogeraki, 2002**] Kalogeraki V., Gunopulos D., Zeinalipour-Yazti D. A Local Search Mechanism for Peer-to-Peer Networks // Proc. of CIKM'02, McLean VA, USA, 2002.

[Landeetc, 2013]Ландэ Д.В., Горбов И.В., Балагура И.В. Характеристики сети соавторов медицинских наук // Клиническая информатика и телемедицина, 2013. - Т.9., Вып.10. - С. 141-144.

[Lande, 2015]Lande D. A Domain Model Created on the Basis of Google Scholar Citations // CEUR Workshop Proceedings (ceur-ws.org). Vol-1536 urn:nbn:de:0074-1536-8. Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015) Obninsk, Russia, October 13-16, 2015.- pp. 57-61.

[Liu, 2015]Liu J., Li Y., Ruan Z., Fu G., Chen X., Sadiq, Deng Y.A new method to construct co-author networks//Phisica A. – 2015. – 419.– P. 29-39.

[Newman, 2003] Newman M.E.J. The structure and function of complex networks // SIAM Rev. – 2003. – 45.– P. 167-256.

[Ortega, 2015]Ortega J. How is an academic social site populated? A demographic study of Google Scholar Citations population// Scientometrics.– 2015. – 104. – P. 1-18

CREATION OF NETWORKS OF THE CO-AUTHORSHIP ACCORDING TO THE GOOGLE SCHOLAR CITATIONS SERVICE

Lande D.V., Balagura I.V., Andrushchenko V.B.

Institute for Information Recording NAS of Ukraine, Kiev , Ukraine

dwlande@gmail.com

balaguraira@mail.ru

valentyna.andrushchenko@gmail.com

The algorithm of creation of the network of a co-authorship of scientists regulated by their scientific interests is given in work. The network of a co-authorship is formed on the basis of sounding of the Google Scholar Citations service. It is shown that the descriptors defining subject influence the size of the formed network, and also dynamics of its growth. It is shown that clusters in networksof a co-authorship can

be considered as a basis for identification of schools of sciences.

Introduction

The objective of the work – is the description of the theoretical principles and methods of automatic formation of the co-authoring networks, in particular in the fields - Complex Networks and Text Mining sounding the great information network. To attain this aim the specific algorithm of Google Scholar Citations service scanning is used to receive the representative co-authors bank as the base units for the future network. Within the sounding notion we will perceive the small size fetch of the most important content from the large networks, which couldn't be sounded by the processing reason [Lande, 2015].

It's evident that the co-authoring network can be of a big size, if is not measured by the defined theme, targeted by the tags of the first author – the origin of the network formation.

Such an effect complicates considerably the perception of the formed network and reduces to the effect of "themes drift". Also the identical last names and initials spelling can occur. To cope with these effects the thematic filtering is used i.e. the used descriptors are referred to authors of the scientometric network, and define their thematic direction.

Accordance of these descriptors in the final analysis defines the size of the formatting co-authoring networks and the dynamic of its growth. In addition the clusters identification in such networks can be perceived as a basis for the science schools, experts' groups etc. extraction [Landeetc, 2013].

Main Part

It is appropriate to use the approved on the peering networks (peer to peer, P2P) models, based on the equality of participants. Peering networks consist of units; each of it interacts only with the several subsets of other units, which corresponds to the co-authoring network.

The sounding of the reference model network is provided according to the next algorithm [Lande, 2015]:

1. The several number of reference (sounding) network units are defined as the basic ones for the new network, according to the sounding results (in the common case the one unit is chosen).
2. For every unit of the reference network the allied units are defined (co-authors), they are added to the network as the result of sounding. The arcs-connections to these units are formed from the root unit.
3. From the current unit of the reference network the pass to the randomly chosen neighboring unit (co-author) is implemented
4. If the circularity takes place or there is the mismatch of the unit to the several measuring condition, the pass to another randomly chosen unit is implemented. If there is no such unit, the network is considered to be built.

Exactly on the results of the quality modeling there was made a conclusion about the opportunity of forming the small branches of connected co-authors, according to the tags, users of Google Scholar Citations service are interested in.

The described algorithm was adapted to the real co-authoring network of Google Scholar Citations in such a way:

1. The first (root) author to begin the sounding is chosen.
2. The list of the basis tags according to the most important conception is defined appraisaly.
3. The page of the web-service of the chosen author opens.
4. All the co-authors from the chosen author profile are added to the forming network.
5. The arcs-connections are tracing to these units (co-authors) from the root unit (author).
6. From the list of the forming network units the unit for the next page transition for the further analysis is chosen randomly. This unit must meet the themes of the chosen subject field (its tags are included into the descriptors, defined on the step 2) and is not the part of the units, which were traced.
7. If such a unit is chosen, so the pass to the step 3 is implemented.
8. If there is no such an author, the network is considered to be built.

According to the described algorithm the process of the network sounding from the several (root) unit is stopped after the circularity, when according to the algorithm the pass is implemented to the unit, been traced, and also if the left units are vary from the main themes (it defines by taking into account the lexical make-up of the tags). And the exact "circularity" is the feature of the pass to another root author or the end of the sounding process.

Conclusion

The suggested attempt is directed to form the networks of co-authorship in frames of the knowledge domain, limited elements of which are the several tags, targeted previously by the scientists – members of the Google Scholar Citations project.

It's necessary to notice that the basic difference of the suggested model of automatic way of the network formation from the existed ones, based on the direct participation of the experts in choosing straight units and connections. In this case the researcher uses only the tiny knowledge parts, inlayed by co-authors, tags, marked as the main for them. Thus the expert environment is widened considerably.

The model is used for the science fields Complex Networks and Text Mining in frames of the Google Scholar Citations, but the suggested attempt can be used for other knowledge domains and scientometric arrays.

The modeling results received using the procedure proposed in the chapter 3 also can be applied to create the subject field model.