САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Математико-Механический факультет

Кафедра системного программирования

Выделение научных сообществ на основе анализа библиографических данных

Курсовая работа студента 361 группы
Ибрагимова Рустама

Научный руководитель:
Суворов Владимир
EMC Санкт-Петербург

Санкт-Петербург
2012

Содержание

1. Введение

За свою историю человечество накопило огромный объём информации в библиотеках, архивах, периодических изданиях и других письменных документах. После преодоления 3-его информационного барьера (создание ЭВМ, [1]) и создания сети Интернет количество информации стало расти с еще большей скоростью, что привело к информационному взрыву [2] и усложнению задачи информационного поиска.

1.1 Электронная библиотека

В конце 20 века во многих странах библиотеки начали создавать электронные версии хранящихся в их фондах материалов, для публикации на веб-ресурсах с открытым доступом – электронных библиотеках.

Примером таких ресурсов являются электронные библиотеки научных статей, которые предоставляют доступ к научным публикациям и собирают текущие публикации по всему миру. На данный момент самыми известными и массовыми электронными библиотеками научных статей являются:

- Digital Library ACM, http://dl.acm.org;
- Google Академия, http://scholar.google.ru;
- CiteSeerX, http://citeseerx.ist.psu.edu;
- arXiv (Cornell University Library), http://arxiv.org.

1.2 Community mining

Community mining, как ответвление Data mining [3], специализируется на извлечении дополнительной, ранее неизвестной, информации из уже существующей, применительно к сообществам людей.

К задачам этой области можно отнести выделение сообществ (community detection) – как одной из самых распространенных и популярных.

К примеру, данные исследования зачастую интересуют работодателей – результаты могут помочь в поиске новых сотрудников. Примером подобного исследования является сервис InMaps http://inmaps.linkedinlabs.com/, анализирующий информацию LinkedIn (http://linkedin.com/, социальная сеть для поиска и установления деловых контактов) для построения карты деловых взаимоотношений.

1.3 Социальный граф

Для Community Mining, как и Data Mining в целом, неотъемлемой частью является наглядное представление результатов анализа. Для визуализации результатов Community mining зачастую используется социальный граф [4].

Социальный граф G = < V, E > - это граф с множеством вершин V, которые представляют людей, и множеством ребер E, представляющих взаимоотношения между людьми. Использование графа обусловлено его легким визуальным восприятием и многообразием алгоритмов работы на нем.

2. Постановка и цели задачи

Современные методы поиска по электронным библиотекам научных статей не позволяют исследователям решать все задачи по поиску информации. В современном мире особенно важно, чтобы исследователи могли следить за трендами в области науки, находить новые темы для исследований, искать схожие своим исследования или исследователей с теми же научными интересами.

Исследуется гипотеза о том, что взаимоотношения между авторами научных статей являются показателем принадлежности данного автора к определенной группе. Рассматриваются следующие взаимоотношения между авторами:

- соавторство,
- цитирование,
- перекрестное цитирование.

Для решения проблемы поиска по электронным библиотекам научных статей предлагается создание эффективного механизма коллаборации ученых на основе библиографических данных с целью выделения научных сообществ.

Основными задачами данного механизма являются:

- выявление научных сообществ,
- поиск научных сообществ близких автору,
- определение позиции автора в глобальном исследовательском обществе и др.

2.1 Обзор существующих решений

Ryutaro Ichise и Hideaki Takeda из Токио уже разрабатывали средства для выделения научных сообществ на базе электронной библиотеки CiNii http://ci.nii.ac.jp. В ходе данной работы была опубликована статья [5], в которой они описали алгоритмы работы своего сервиса. (см. рис. 1)

Отчасти именно работа Takeda и Ichise стала мотивацией для данного исследования, т.к. разработанный ими сервис доступен лишь на японском языке, а с момента публикации [5] новостей про данное исследование найти не удалось.

В ходе данной работы были изучены публикации об алгоритмах выделения групп на графах [6, 7, 8], в которых рассматриваются различные алгоритмы: для работы на очень больших графов, для сложных графов и др.

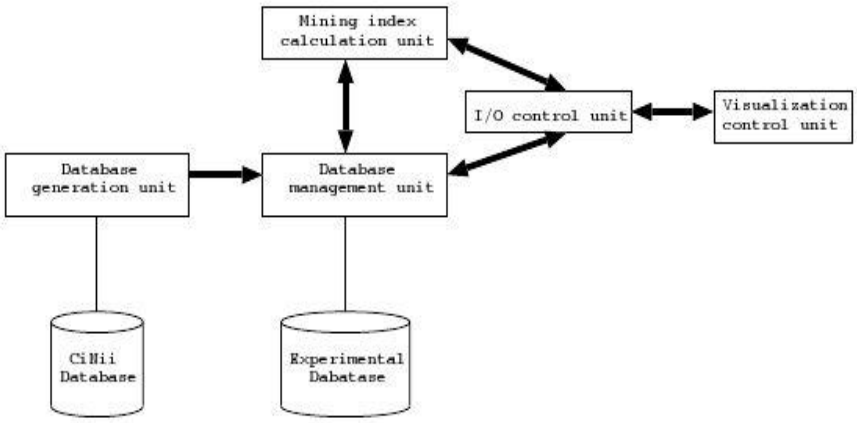

Рисунок 1. CiNii community mining tool

3. Реализация

3.1 Этапы реализации

Решение поставленной задачи осуществлялось в несколько этапов:

1) Автоматизированный сбор контента
2) Унификация данных
3) Создание единой базы данных и ее наполнение
4) Алгоритм выделения групп
5) Визуализация

3.2 Автоматизированный сбор контента

Данные для последующей работы были получены на основе электронных библиотек Digital Library ACM и CiteSeerX.

Для этого были реализованы парсеры (синтаксические анализаторы) на языке программирования Java, автоматизирующие процесс сбора и выделения необходимых библиографических данных авторов научных статей.

В связи с ограничениями сайта Digital Library ACM – блокирование пользователей за частые обращения к ресурсу – процесс сбора контента с этого сайта был распараллелен на несколько компьютеров (создавались локальные копии страниц ресурса для последующей работы парсера с этими данными).

В целях увеличения скорости работы парсера он был реализован для многопоточного выполнения:

- `Thread1` создавал очередь и загружал данные для парсинга.
- `Thread2` и `Thread3` параллельно выполняли парсинг данных.

3.3 Унификация данных

Так как данные с двух электронных библиотек предполагалось использовать совместно, они были приведены к единообразному виду - использовались библиографические данные, представленные на обоих ресурсах.

Использовались следующие данные для авторов:

- имя и фамилия автора,
- место работы автора,
- уникальный идентификационный номер автора на ресурсе,

- список публикаций автора.

Использовались следующие данные для научных статей:

- название статьи,
- авторы,
- краткая аннотация,
- уникальный идентификационный номер статьи на ресурсе,
- издатель,
- год публикации.

3.4 Создание единой базы данных и ее наполнение

Для хранения библиографических данных, полученных в ходе работы парсера, использовалась СУБД MySQL. Данная СУБД распространяется под лицензией GNU GPL и является наиболее удобной для малых и средних баз данных.

ER-диаграмма базы данных изображена на рис. 2.

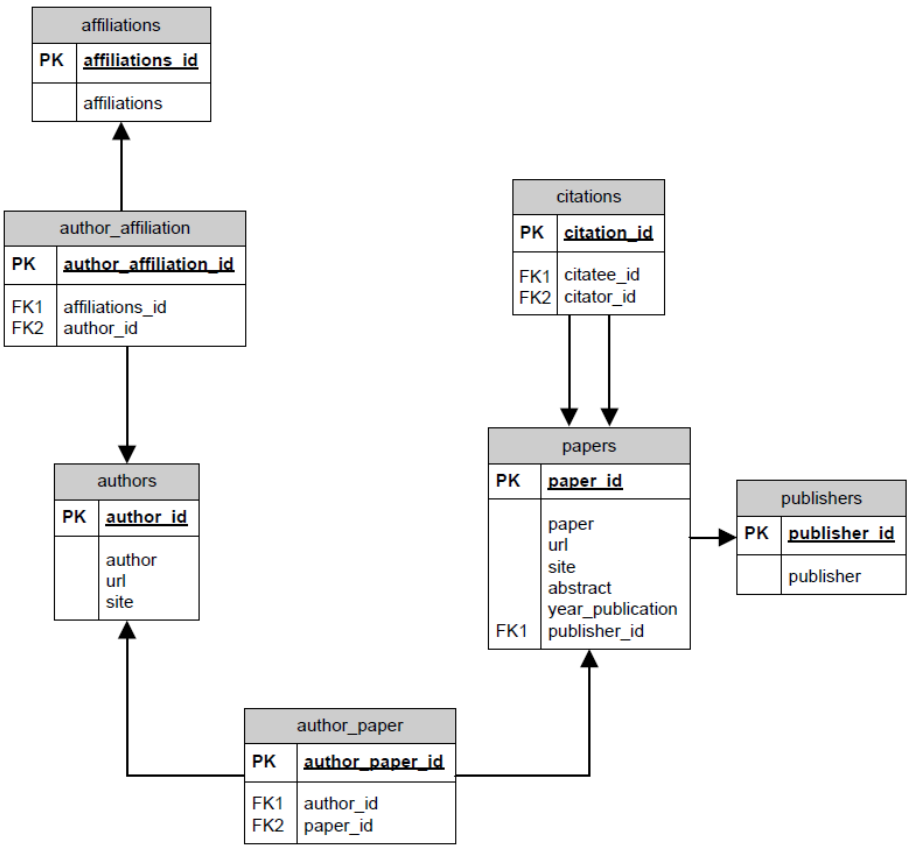

Рисунок 2. ER-диаграмма базы данных.

Для улучшения быстродействия базы данных были введены индексы для полей `papers:paper` и `authors:author`, т.к. они будут использоваться при выполнении запросов пользователей.

Наполнение базы данных производилось во время работы Java-парсера. Количественные показатели единой базы данных приведены в таблице 1.

	Digital Library ACM	*CiteSeerX*	*Общее*
Авторы	307126	16154	323280
Статьи	359188	199462	558650
Количество ссылок	1493332	459883	1953215

Таблица 1. Количественные показатели единой базы данных

3.5 Алгоритм выделения групп

Алгоритм, используемый в данной работе, основан на “Bron-Kerbosch algorithm with pivot vertex” [9]. В данном алгоритме используется 3 множества вершин R, P и X, он находит максимальные клики, которые содержат вершины из R, некоторые вершины из P и не содержат вершин из X.

Дополнение алгоритма “with pivot vertex u from set P” было сделано авторами данного алгоритма для уменьшения количества рекурсивных вызовов алгоритма, так как максимальная клика содержит либо вершину `u`, либо любую вершину из множества `P \ N(u)` (где `N(u)` – множество вершин, смежных с `u`). Тем самым отсекались ветви решения с изначально не максимальными кликами. Позже, в работе [10], было доказано, что вершину `u` необходимо выбирать из множества `P ∪ X`.

Ниже приведен псевдокод алгоритма с дополнением “with pivot vertex” и дополнением из работы [10].

```
BronKerbosch(R,P,X):
  if P and X are both empty:
    report R as a maximal clique
  choose a pivot vertex u in P ⋃ X
  for each vertex v in P \ N(u):
    BronKerbosch(R ⋃ {v}, P ⋂ N(v), X ⋂ N(v))
    P := P \ {v}
    X := X ⋃ {v}
```

На входе алгоритма граф G = < V, E >, где V – это множество вершин-авторов, полученных из базы данных по поисковому запросу пользователя, E – множество ребер, обозначающих связи между этими авторами.

Результат работы алгоритма – это граф с выделенными другим цветом максимальными кликами. Дополнительно другим цветом выделены вершины, смежные хотя бы с одной вершиной из клики. В контексте социального графа это авторы, не входящие в научное сообщество, но связанные с ним.

3.6 Визуализация

Визуализация результатов осуществлялась с помощью социального графа, который строился с использованием Jung Framework. Пример социального графа приведен на рис. 2.

В социальном графе реализованы 2 типа ребер:

- цитирование – синий цвет,
- соавторство – красный цвет.

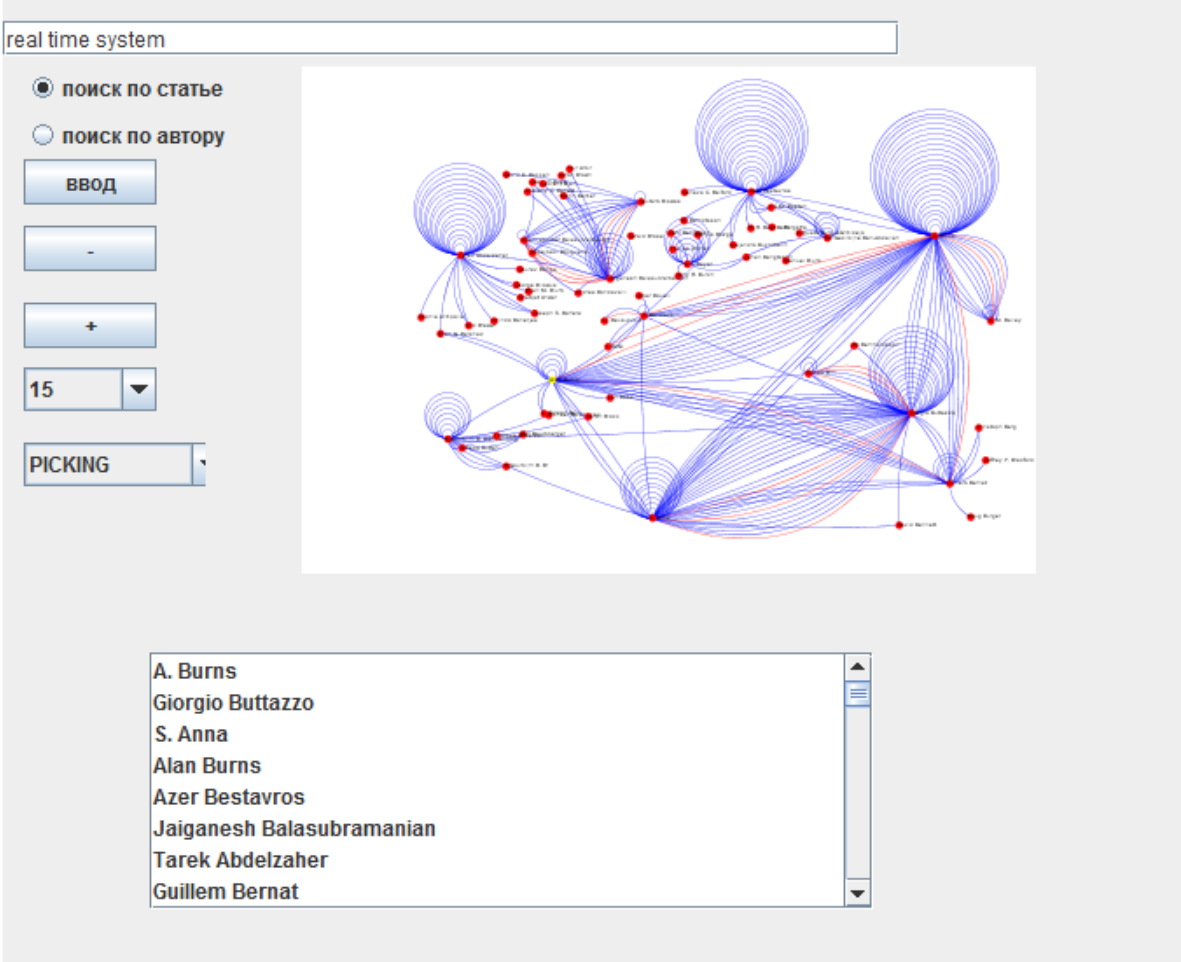

Рисунок 2. Интерфейс приложения

3.7 Веб-сервис

Результат работы был оформлен в виде веб-сервиса, реализованного с использованием Java-апплета и Java-сервлета по принципу клиент-серверного приложения.

Преимущества Java-апплетов:

- кроссплатформенность,
- работа на всех версиях Java,
- система сервер/клиент.

Веб-сервис позволяет пользователю сформировать запрос поиска по автору или по статье и визуализирует результаты запроса в виде социального графа.

Обработкой запроса пользователя, работой с базой данных и построением графа занимается сервлет. Апплет работает на стороне клиента и занимается визуализацией.

4. Результаты

В ходе исследования создан прототип приложения, использующий базы электронных библиотек Digital Library ACM и CiteSeerX, и позволяющий по параметрам пользователя осуществлять поиск и визуализировать научные сообщества с помощью социального графа.

Были изучены существующие решения для community mining в области научных публикаций и алгоритмы выделения групп на графах.

Дальнейшие перспективы данной работы – это создание полноценного веб-сервиса для выделения научных сообществ, с разработкой собственного алгоритма работы на графе и улучшение визуализации результатов выделения научных сообществ.

Ссылки

[1] http://ru.wikipedia.org/wiki/Информационный_барьер

[2] http://en.wikipedia.org/wiki/Information_explosion

[3] http://en.wikipedia.org/wiki/Data_Mining

[4] http://en.wikipedia.org/wiki/Social_graph

[5] Ryutaro Ichise , and Hideaki Takeda. Community mining tool using bibliography data. Proceedings of the 9th International Conference on Information Visualization, 2005.

[6] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks, 2004.

[7] Nan Du, Bin Wu, and Bai Wang. Community Detection in Complex Networks, Journal of Computer Science and Technology, Volume 23 Issue 4, 2008

[8] Bo Yang, William K. Cheung, and Jiming Liu. Community mining from Signed Social Networks, IEEE Transactions on knowledge and data engineering, vol. 19, no. 10, 2007

[9] Coen Bron, and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. Communications of the ACM, Volume 16 Issue 9, Sept. 1973

[10] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science Volume 363, Issue 1, 25 October 2006