НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК АЗЕРБАЙДЖАНА
ИНСТИТУТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

На правах рукописи

ФАДАИ САРАФРАЗ оглу ГЯНДЖАЛИЕВ

РАЗРАБОТКА МЕТОДОВ И АЛГОРИТМОВ ДЛЯ ВЫЯВЛЕНИЯ И АНАЛИЗА НАУЧНЫХ СОЦИАЛЬНЫХ СЕТЕЙ НА ВЕБ

3338.01 – Системный анализ, управление и обработка информации

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени
доктора философии по технике

Баку – 2013

Диссертационная работа выполнена в Институте Информационных Технологий Национальной Академии Наук Азербайджана

Научный руководитель:

доктор технических наук **Р.М.Алыгулиев**

Официальные оппоненты:

доктор технических наук, доцент **В. Г. Мусаев**

доктор философии по технике **Л. Э. Керимова**

Ведущая организация: Бакинский Государственный Университет (кафедра «Информационных технологий и программирования»)

Защита состоится **19 июня 2013 года в 14^{00}** часов на заседании диссертационного совета FD.01.231 при Институте Информационных Технологий НАН Азербайджана. **Адрес**: AZ1141, г. Баку, ул. Б. Вахабзаде, 9.

С диссертацией можно ознакомиться в библиотеке Института Информационных Технологий НАН Азербайджана.

Автореферат разослан **« 16 » мая 2013 г**.

Ученый секретарь диссертационного
совета, доктор философии по технике **Р.Г.Шыхалиев**

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Во второй половине 90-х годов прошлого века происходит экспоненциальное развитие сети Интернет и связанных с ним информационно-коммуникационных технологий. В конце XX века в Интернете насчитывалось более 300 млн. постоянно подключенных к нему серверов. Интернет занимает все большее место в жизни современного общества, растет количество, как самих сайтов различных назначений, так и пользующихся этими сайтами пользователей.

Анализ социальных сетей является достаточно недавно возникшей дисциплиной, которая начала приобретать известность несколько десятилетий назад. Перед тем, как это направление начало изучаться широко, с анализом социальных сетей можно было столкнуться лишь в некоторых работах исследователей. Однако, с момента появления Веб, анализ социальных сетей приобретал все большую известность и начал привлекать внимание исследователей. В основном, причиной этому служило то, что до появления Веб существовало всего лишь несколько способов сбора данных для создания социальных сетей и извлекаемые социальные сети были малых размеров.

Таким образом, Веб является источником данных, откуда можно извлечь информацию для выявления отношений между различного рода ее участниками. Ученые, научные исследователи и научные институты не являются исключениями в этом смысле. На Веб сайтах научных журналов, исследователей, цифровых библиотек можно приобрести обширную информацию о деятельности научных исследователей в той или иной области. С помощью этой информации можно проанализировать прошлую и настоящую и предвидеть будущую деятельность исследователей.

Современная наука обладает множествами направлений и в каждом из них существует большое количество нерешенных задач. Это, в свою очередь, побуждает ученых и научные институты к групповому исследованию. Для развития и поддержания группового научного исследования, определения приоритетных направлений, определения ведущих научных исследователей, оценки их научной деятельности социальная сеть научных исследователей и институтов играет важную роль. Становится ясным, что по сравнению с другими социальными сетями социальная сеть научных исследователей и институтов отражает более точную, богатую и надежную информацию. Для анализа и оценки научной деятельности исследователей и институтов, созданных на основе списка научных работ, ссылок, соавторства, участия в конференциях, деятельности в научном журнале выявления и

анализ социальных сетей научного сообщества является актуальной проблемой на данный момент.

В результате усилий исследователей были созданы методы и алгоритмы для выявления и анализа социальных сетей на Веб. Количество таких работ небольшое. Некоторая часть из них применима к социальных сетям с ограниченным количеством акторов. Другая часть, ограничивает количество связей, которые могут быть между акторами. Кроме того, некоторые подходы применимы только в конкретных областях и показывают хорошие результаты только при определенных условиях.

Цель и задачи работы. Целью диссертационной работы является разработка новых методов и алгоритмов для выявления и анализа социальных сетей научного сообщества на Веб. В соответствии с указанной целью в работе поставлены следующие основные научные задачи:

- разработка методов для выявления и синтеза социальных сетей научных исследователей на Веб;
- разработка метода для выявления социальных сетей научных учреждений на Веб;
- исследование роли меры близости и алгоритма ранжирования в анализе социальных сетей, извлеченных на Веб;
- разработка метода и алгоритма для выявления сообществ в социальных сетях, извлеченных на Веб;

Методы исследования. При решении поставленных задач использовались теория матриц, теория графов, теория нечетких множеств, целочисленное программирование, генетические алгоритмы, спектральный анализ.

Основные положения, выносимые на защиту:

- методы для выявления и синтеза социальных сетей научных исследователей на Веб;
- метод для выявления социальных сетей научных учреждений на Веб;
- роль меры близости и алгоритмов ранжирования в анализе социальных сетей, извлеченных на Веб;
- метод и алгоритм извлечения сообществ из социальных сетей, построенных на основе информации, полученной на Веб.

Научная новизна. В рамках диссертационной работы были получены следующие основные результаты, обладающие научной новизной:

- предложена аналитическая модель для синтеза социальных сетей научных исследователей, извлеченных из нескольких источников данных на Веб;

- предложен метод выявления социальной сети научных учреждений на основе информации, доступной на Веб;
- проведен анализ роли меры близости и алгоритма ранжирования в анализе социальных сетей, извлеченных на Веб;
- предложен метод извлечения сообществ из социальных сетей, полученных на Веб, как задача целочисленного программирования; разработан генетический алгоритм для ее решения.

Практическая ценность работы. Полученные научно-теоретические и практические результаты могут быть использованы при:

- принятии эффективных решений в сфере э-государства;
- интеллектуальном анализе в различных научных направлениях;
- принятии решений в сфере э-науки;
- управлении персоналом в предприятиях;
- разработке поисковых систем в интернете;

Апробация работы. Основные научно-теоретические и практические результаты докладывались и обсуждались на: научной конференции докторантов Национальной Академии Наук Азербайджана (Баку, Азербайджан, май, 2010 г.); Third International Conference «Problems of Cybernetics and Informatics» – PCI'2010 (Baku, Azerbaijan, september, 2010); научной конференции докторантов Национальной Академии Наук Азербайджана (Баку, Азербайджан, май, 2011 г.); Fourth International Conference «Problems of Cybernetics and Informatics» – PCI'2012 (Baku, Azerbaijan, september, 2012); научном семинаре Института Информационных Технологий Национальной Академии Наук Азербайджана (Баку, Азербайджан, март, 2013 г.).

Публикации. По теме диссертации опубликованы 10 печатных работ, в том числе 6 статей в рецензируемых международных научных журналах и 4 тезиса.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы из 139 названий и приложения. Общий объем работы – 147 страниц, основной текст – 127 страниц. В работе имеются 24 таблиц и 38 рисунков.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи, приведены основные положения, выносимые на защиту, определены научная новизна и практическая ценность полученных результатов.

Первая глава посвящена проблеме выявления и анализа социальных сетей на Веб. В этой главе анализируется актуальность извлечения информации об отношениях между акторами и последующее построение социальных сетей на Веб.

В первую очередь, дается подробная история возникновения дисциплины анализа социальных сетей, и описываются работы, которые сыграли ключевые роли в приходе на сцену этого направления. Затем, описываются и подробно расказывается об основных понятиях и методах анализа социальных сетей.

Следующим шагом, рассказывается о том, как приход на сцену Веб повлиял на анализ социальных сетей и каким образом были пересмотрены методы анализа социальных сетей в связи с этим. Затем, анализируются работы, посвященные выявлению и анализу социальных сетей, извлеченных на основе информации полученной из разных источников данных на Веб. Вначале, рассказывается о разнородных источниках информации для выявления социальных сетей, таких как электронная почта, журнальные файлы, новостные статьи и т.д. Далее, рассматриваются работы, посященные выявлению и анализу социальных сетей научного сообщества. Выявляются преимущества и недостатки перечисленных работ, а также говорится о некоторых програмных обеспечениях, основанных на подобных работах и случаев их применения.

Кроме этого, проводится обширный анализ современных методов и алгоритмов выявления сообществ в социальных сетях на Веб. Выявляются преимущества и недостатки этих работ. Также подробно описываются работы, сыгравшие ключевые роли в этом направлении.

В последней части первой главы выделяются некоторые нерешенные проблемы в области выявления научных социальных сетей и выявления сообществ, а также обосновывается актуальность работы. Эти проблемы группируются как поставленные задачи в диссертационной работе, конкретизируются методы и алгоритмы, которые необходимо разработать и перечисляются применяемые подходы при решении поставленных задач.

Во второй главе рассказывается о методах выявления, приводятся методы для синтеза социальных сетей научных исследователей, и предлагается метод выявления социальных сетей научных учреждений.

Научные исследователи могут иметь отношения разных видов, такие как, со-цитирование, участие в проектах, соавторство и т.д. Поэтому первая задача ставится следующим образом. Предполагается, что имеются несколько источников данных, откуда можно извлечь информацию об отно-

шениях некоторого количества заранее заданных научных исследователей. Каждый из этих источников данных отличается от остальных по виду содержащих отношений. Следовательно, каждый из этих источников информации будет иметь разный степень доверия. Задачей является слияние социальных сетей, выявленных из этих источников в одну результирующую социальную сеть с минимальной потерей информации и так, чтобы можно было присваивать коэффициент важности для каждого источника.

Для синтеза нескольких социальных сетей научных исследователей вводится следующая формула:

$$w_{ij} = \sum_{k=1}^{m} \widetilde{w}_k w_{ij}^k \ , k = 1,...,m, \tag{1}$$

где m - количество первоначальных социальных сетей, w_{ij} – вес связи акторов i и j результирующей социальной сети, $\widetilde{w}_k \in [0,1]$, $k = 1,...,m$ – коэффициент, присваиваемый источнику k и w_{ij}^k – вес связи акторов i и j в сети k, $i, j = 1,..., n$. Переменные $\widetilde{w}_k \in [0,1]$ выбираются на основе наблюдений или экспериментальным путем, для которых $\sum_{k=1}^{n} \widetilde{w}_k = 1$.

Первый подход для нахождения неизвестных коэффициентов в формуле (1) подразумевает рассмотрение этой формулы как OWA-оператор Ягера. Оператор Ягера определяется так:

$$F : R^n \to R,$$

который имеет связанный вектор весов $\widetilde{W} = (\widetilde{w}_1, \widetilde{w}_2, ..., \widetilde{w}_n)$ такой, что $\widetilde{w}_i \in [0,1]$, $1 \le i \le n$ и $\sum_{i=1}^{n} \widetilde{w}_i = 1$. Также,

$$F(a_1, a_2, ..., a_n) = \widetilde{w}_1 b_1 + \widetilde{w}_2 b_2 + ... + \widetilde{w}_n b_n, \tag{2}$$

где b_j, $j = 1,..., n$, равен значению наибольшего элемента во множестве $< a_1, a_2, ..., a_n >$. Элемент a_i не связан с определенным весом w_i, а вес w_i связан с определенным порядковым положением элемента a_i.

Для нахождения неизвестных коэффициентов в формуле (2) используется один из предложенных методов в литературе, который требует решения следующего полиномиального уравнения:

$$\widetilde{w}_1 \left[(n-1)\alpha + 1 - n\,\widetilde{w}_1\right]^n = ((n-1)\alpha)^{n-1} \left[((n-1)\alpha - n)\widetilde{w}_1 + 1\right] \tag{3}$$

и

$$\widetilde{w}_n = \frac{((n-1)\alpha - n)\widetilde{w}_1 + 1}{(n-1)\alpha + 1 - n\widetilde{w}_1}. \quad (4)$$

Здесь α вычисляется следующим образом:

$$\frac{1}{n-1}\sum_{i=1}^{n}(n-i)\widetilde{w}_i = \alpha, \quad (5)$$

где $0 \le \alpha \le 1$, $i = 1,...,n$ и называется мерой *orness*, которая классифицирует OWA-оператор по отношению нахождения между "and" и "or"(0 и 1). Чем больше значение *orness*, тем больше принимающий решение "идет на риск".

При втором подходе задача синтеза нескольких социальных сетей рассматривается как задача многокритериального выбора на основе нечеткого отношения предпочтения, суть которого заключается в следующем.

Предположим, что имеется n альтернатив $S = (s_1, s_2, ..., s_n)$. Также, допустим, что каждая из этих альтернатив обладает m критериями $C = (c_1, c_2, ..., c_m)$. Предполагается, что известна информация о попарном сравнении каждой из альтернатив по каждому из критериев. В теории многокритериального выбора, каждый критерий представляется в виде следующего нечеткого множества:

$$c_j = \left\{ \frac{w_1^{(j)}}{s_1}, \frac{w_2^{(j)}}{s_2}, ..., \frac{w_n^{(j)}}{s_n} \right\}, \; j = 1,2,...,m. \quad (6)$$

Числа $w_i^{(j)}$, показывающие "вклад" альтернативы s_i в критерий c_j принимают значения из отрезка [0,1]. Эти числа могут представлять собой весы альтернатив по каждому из критериев. Для этих чисел выполняется следующее условие:

$$w_1^{(j)} + w_2^{(j)} + ... + w_n^{(j)} = 1, \; j = 1,2,...,m. \quad (7)$$

Согласно принципу Беллмана-Заде лучшую альтернативу s^* следует искать в пересечении нечетких множеств, представляющих критерии, т.е.

$$s^* \in D = c_1 \bigcap c_2 \bigcap ... \bigcap c_m. \quad (8)$$

В теории нечетких множеств задачу нахождения решения из пересечения множеств можно привести к задаче минимизации, т.е. справедливо $\cap \to \min$. Поэтому множество потенциальных решений приобретает следующий вид:

$$D = \left\{ \frac{\min\left\{w_1^{(1)}, ..., w_1^{(m)}\right\}}{s_1}, \frac{\min\left\{w_2^{(1)}, ..., w_2^{(m)}\right\}}{s_2}, ..., \frac{\min\left\{w_n^{(1)}, ..., w_n^{(m)}\right\}}{s_n} \right\}. \quad (9)$$

Таким образом, в качестве лучшей альтернативы необходимо искать альтернативу с максимальным весом:

$$w(s^*) = \max_{i=1,2,...,n} \min\left\{w_i^{(1)}, w_i^{(2)}, ..., w_i^{(n)}\right\}. \quad (10)$$

Для нахождения весов альтернатив в формуле (7) существует метод наихудшего случая:

$$w_1 = r_1 \frac{w_l}{r_l}, w_2 = r_2 \frac{w_l}{r_l}, ..., w_n = r_n \frac{w_l}{r_l} \ , \quad (11)$$

где r_i $(i = 1,2,...,n)$ ранг альтернативы s_i с весом w_i и w_l вес наихудшей альтернативы s_l по критерию c_j $(j = 1,2,...,n)$ и рангом r_l. Из формул (7) и (11) вытекает, что:

$$w_l = \frac{1}{\frac{r_1}{r_l} + \frac{r_2}{r_l} + ... + \frac{r_n}{r_l}} = \frac{1}{\sum_{i=1} \frac{r_i}{r_l}} . \quad (12)$$

Здесь, соотношение $\frac{r_i}{r_j}$ берется из 9-ти бальной шкалы Саати, в которой это отношение равно одному из чисел в интервале [1, 9], в зависимости от того, насколько альтернатива s_i лучше альтернативы s_j по данному критерию. Из вышеописанного видно, что для определения весов альтернатив необходимо знать какая альтернатива является наихудшей по каждому из критериев. Для социальных сетей выбираются несколько критериев, социальные сети рассматриваются как альтернативы и проводится слияние этих сетей с помощью описанного подхода.

Во второй части второй главы предлагается метод, с помощью которого можно получить социальную сеть научных учреждений. Поиск среди существующих методов выявления социальных сетей на Веб, показал, что до сих пор не была адресована задача выявления социальной сети научных учреждений, с помощью информации, доступной на Веб.

Основывасяь следующие три вида связей между научными учреждениями предлагается метод выявления их социальных сетей:

1. Научные учреждения публикуют результаты своих работ в виде научных статьей в научных журналах, которые обычно размещаются на сайтах этих учреждений и в некоторых цифровых библиотек.

2. Научные учреждения периодически проводят семинары и конференции по определенным темам; члены научных учреждений могут быть спикерами собрания.
3. Члены разных научных учреждений могут быть также членами редколлегии одного и того же научного журнала.

Третья глава посвящена исследованию влияния меры близости на ранжирование акторов и выявлению сообществ в социальных сетях.

Так, в первой части этой главы описывается один из первых алгоритмов ранжирования веб-страниц PageRank и рассказывается о различных его модификациях. Алгоритм ранжирования веб-страниц переводится на социальные сети с помощью того, что веб-страницы рассматриваются как акторы сети, а гиперссылки между страницами рассматриваются как связи между акторами. Становится видно, что большинство методов ранжирования в той или иной мере зависят от меры подобия акторов. В этой части третьей главы выдвигается предположение, что в социальных сетях выбранная мера подобия между акторами влияет на ранжирование акторов и если влияет, то проанализировать каким образом.

Во второй части третьей главы предлагается и детально описывается метод для выявления сообществ в социальных сетях. Метод основывается на максимизации предлагаемой целевой функции. Предложенный метод относится к группе методов *партиционированной кластеризации*, поскольку кластеризация является одним из способов выявления сообществ в социальных сетях. Принято, что задача кластеризации ставится следующим образом.

Допустим, имеется набор акторов $O = \{o_1, o_2, ..., o_n\}$ социальной сети. Процесс кластеризации представляет собой разбиение набора акторов O на некоторое количество k непересекающихся групп $C_1, C_2, ..., C_k$. Группы, полученные в результате разбиения, называют *кластерами*. При этом, у двух разных кластеров общих акторов быть не должно, не должно быть акторов не принадлежащих к какому-либо кластеру и каждый кластер должен содержать хотя бы одного актора.

Проблема выявления сообществ формулируется как следующая задача целочисленного программирования ((13) – (17)):

$$f(x,y) = \beta_1 \sum_{q=1}^{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e_i w_{ij} x_{iq} x_{jq} y_q - \beta_2 \cdot \sum_{p=1}^{n-1} \sum_{q=p+1}^{n} w_{pq} y_p y_q \to \max, \qquad (13)$$

$$\sum_{q=1}^{n} x_{iq} = 1,\ i = 1,2,...,n, \tag{14}$$

$$\sum_{i=1}^{n} x_{iq} \geq 1,\ q = 1,2,...,n, \tag{15}$$

$$\sum_{q=1}^{n} y_q = k,\ q = 1,2,...,n, \tag{16}$$

$$x_{iq} \leq y_q,\ \ i,q = 1,2,...,n. \tag{17}$$

Здесь, первая величина в формуле (13) есть сумма весов в выбранных кластерах, а вторая величина определяет близость между кластерами. Также $\beta_1 + \beta_2 = 1$, которые являются относительным вкладом в каждое из двух величин в формуле (13) соответственно. Значения этих величин выбираются экспериментальным путем или путем наблюдений, а e_i, $i = 1,2,...,n$ есть степень вершины o_i. Условие (14) означает, что каждый актор должен быть членом только одного кластера. Условие (15) означает, что каждый кластер должен быть непустым и должен содержать хотя бы одного актора. Условие (16) означает, что необходимо гарантировать, что выбрано только k кластеров. Условие (17) означает, что кластер выбран, если имеется какой-либо актор, присвоенный этому кластеру.

Для максимизации данной целевой функции предлагается использование *генетического алгоритма (ГА)*. Генетический алгоритм входит в класс эвристических алгоритмов. Преимуществами этого алгоритма являются: простота использования, большая вероятность нахождения приблизительного решения случайным образом, меньшее количество ограничений, возможность изменения решения на лету и т.д.

В предлагаемом методе количество кластеров должно быть заранее задано. После этого, для каждой возможной комбинации пары акторов (как центроидов), необходимо вычислять значение функции (13), пытаясь достичь максимально возможного. Например, если количество кластеров равно двум, то необходимо для данного значения вектора x вычислить значение функции для всех возможной комбинаций y_{ij}, где $i, j = \overline{1,n}$. Когда, акторы o_i и o_j выбраны как центры кластеров, то $y_i = 1$ и $y_j = 1$ и $y_p = 0$, где $p = 1,n$, $p \neq i, j$. Для больших значений n такой объем вычислений может потребовать большие вычислительные ресурсы. Для избежания подобного, необходимо вычислять значение целевой функции только для тех пар o_i и o_j, вы-

бранных как центры кластеров, которые могут быть отобраны как кандидаты для этого. Следовательно, чтобы избежать лишних вычислений, необходимо воспользоваться каким-либо критерием отбора.

В качестве такого критерия воспользуемся *степенью информированности* актора. Эта величина показывает, какой объем информации содержит актор и вычисляется с помощью формулы:

$$\alpha_i = IC(i) = \frac{n}{nc_{ii} + (T - 2R_i)}, \tag{18}$$

где $T = \sum_{j=1}^{n} c_{jj}$, $R_i = \sum_{j=1}^{n} c_{ij}$, c_{ij} $(i = 1,2,...,n)$ элементы матрицы $C = B^{-1}$, $B = D - A + I$, D – диагональная матрица с элементами $d_{ii} = \sum_{j=1}^{n} a_{ij}$ $(i = 1,2,...,n)$, I – единичная матрица и A – матрица, элементами a_{ij} которого являются значения весов, соединяющих соответствующие вершины i и j. Чем больше значение степени информированности актора, тем большей информацией и влиянием он обладает и, следовательно, тем больше он подходит как центр кластера.

В качестве первоначальной популяции создается группа подобных векторов-индивидов со случайным распределением акторов между кластерами. Далее, в каждом поколении каждая особь скрещивается с каждым другим методом одноточечного скрещивания. Идентификация приспособленности (fitness) заключается в сравнении текущего значения функции с предыдущим значением. Если обнаруживается, что значение функции увеличивается, что делаем вывод, что текущий индивид является более приспособленным и “лучше выживет” в следующем поколении кандидатов. В этом случае индивид также включается в следующее поколение, если при нем функция принимает большее значение. В данном случае мутация не используется. Скорость сходимости функции $O(n^2 p)$, где p – количество акторов с наибольшим значением степени информированности, выбранных как кандидаты на центры кластеров.

В четвертой главе проводится анализ предложенного метода объединения социальных сетей научных исследователей, анализируется роль меры близости в ранжировании акторов и проводится эксперимент для проверки предложенного метода выявления сообществ.

Интересным фактом в формуле (1) является то, что информация о существовании связи никогда не будет утеряна. Вес связи в результатирующей социальной сети будет равен нулю только в том случае, когда в каждой из составляющих сетей весы соответствующих связей будут равны нулю одновременно.

Для проверки метода слияния социальных сетей научных исследователей используются произвольно взятые три социальные сети из пяти акторов (n=5, m=3). В каждом из этих социальных сетей между акторами имеется определенный вес. В следующей таблице показаны значения выбранных критериев для каждого из сетей (для метода 2):

Таблица 1

Значения критериев социальных сетей, полученных из отдельных источников

Сеть	Значение критерия		
	Средний вес связи	Плотность	Среднее расстояние
1	0,225	0,660	0,460
2	0,200	0,480	0,460
3	0,342	0,985	0,712

В следующей таблице (таблица 2) показаны значения критериев для результирующей социальной сети, полученной с помощью метода 2:

Таблица 2

Значения критериев для результатирующей многородной социальной сети

средний вес связи	плотность	среднее расстояние
0,2940	0,4152	0,3019

Из таблицы 1 видно, что для каждой социальной сети значение плотности больше значения веса и расстояния. Это совпадает со сравнением в таблице 2 для результирующей сети. Другими, словами при “переходе” от первоначальных социальных сетей к результирующему, также передаются свойства сетей к результирующему.

Для дополнительного анализа методов слияния социальных сетей научных исследователей выбирается какой-либо показатель актора, например, *степень вершины*. Далее, вычисляются значения этого параметра в начальных сетях и полученных результирующих сетях. В следующей таблице (таблица 3) акторы отсортированы согласно значениям их степени вершины в однородных сетях в убывающем порядке:

Таблица 3

Ранжирование акторов по значению степени вершины в однородных социальных сетях

Ранг	Акторы в сетях		
	1	2	3
1	a_5	a_5	a_5
2	a_4	a_4	a_4
3	a_1	a_1	a_3
4	a_2	a_2	a_1
5	a_3	a_3	a_2

Отсортируем таким же образом акторов в результирующей социальной сети за счет весов, полученных с помощью метода 1:

Таблица 4

Ранжирование акторов по значению степени вершины методом 1

Ранг	Акторы в сетях		
	$\alpha = 0{,}2$	$\alpha = 0{,}5$	$\alpha = 0{,}8$
1	a_5	a_5	a_5
2	a_4	a_4	a_4
3	a_1	a_1	a_3
4	a_2	a_2	a_1
5	a_3	a_3	a_2

Отсортируем таким же образом акторов в результирующей социальной сети за счет весов, полученных с помощью метода 2:

Таблица 5

Ранжирование акторов по значению степени вершины методом 2

Ранг	Актор
1	a_5
2	a_4
3	a_3
4	a_2
5	a_1

Из таблицы 3 видно, что акторы a_5 и a_4 во всех трех однородных социальных сетях имеют самые высокие ранги, а именно 1 и 2, соответственно. В синтезированных социальных сетях, полученных обоими методами (таблицы 4 и 5), эти акторы обладают такими же рангами. Далее, в методе 1 при $\alpha = 0{,}2$ и $\alpha = 0{,}5$ ранжирование акторов в однородных социальных сетях полностью совпадает с ранжированием акторов в результирующей сети, полученного с помощью метода 1. При увеличении значении α ранг актора a_1 ухудшается: понижается с 3-го на 4-ый. Также ухудшается ранг актора a_2, понижаясь с 4-го на 5-ый. При увеличении значения α ранг актора a_3 наоборот увеличивается, поднимаясь с 3-го на 5-ый. Кроме этого, при $\alpha = 0{,}8$ ранжирование акторов полностью совпадает с третьей из однородных сетей. Чем больше значение α, тем больше мы "готовы рискнуть" с помощью метода, на основе которого принимаем решение. В методе 2 можно сказать, что результирующая социальная сеть совпадает с третьей из однородных социальных сетей, за исключением того, что акторы a_1 и a_2 меняются рангами.

Для анализа влияния меры близости на ранжирование акторов пользуемся следующим графом:

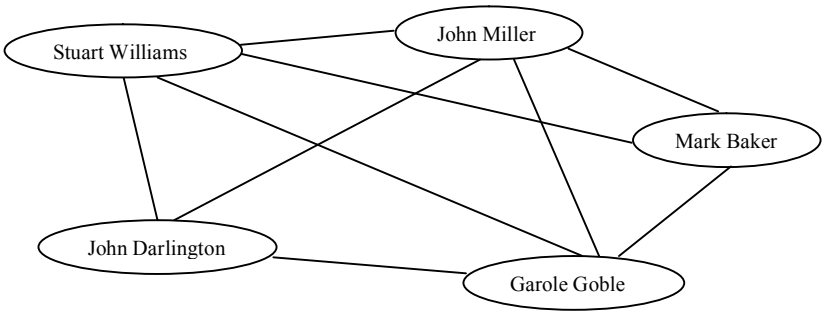

Рис. 1. Социальная сеть пятерых участников конференции WWW2006.

Обозначим акторов через e_1, e_2, e_3, e_4, e_5 начиная со Stuart Williams по направлению часовой стрелки на графе. Первым шагом вычисляются весы связей с помощью трех мер подобий: коэффициента Жаккара, коэффициента перекрытия и нормализованного Google расстояния (*Normalized Google Distance* – NGD). С помощью этих весов акторы ранжируются использованием алгоритма Topic-Centric:

$$PR(e_j) = \frac{(1-d)}{n} + d \sum_{v_i \in B(v_j)} \frac{sim(e_i, e_j)}{\sum_{v_k \in F(v_i)} sim(e_i, e_k)} PR(v_i), \quad (19)$$

где $sim(e_i, e_j)$ – мера подобия/близости акторов e_i и e_j, $B(e_i)$ – множество акторов, которое ссылается на актор e_i, а $F(e_i)$ – множество акторов, на которое ссылается актор e_i, $d \in [0.8,1]$ – коэффициент смягчения. Затем, с помощью формулы

$$sim(e_i, e_j) = \frac{1}{r(e_i, e_j)}, \quad (20)$$

где $r(e_i, e_j)$ – расстояние сопротивления между акторами e_i и e_j, вычисляются новые весы между акторами и с помощью того же алгоритма акторы заново ранжируются. Результаты ранжирования показаны в следующей таблице:

Таблица 6

Результаты ранжирования акторов первоначальных и производных социальных сетей

Актор	Социальная сеть Жаккара		Социальная сеть перекрытия		Социальная сеть NGD	
	$Ранг^{TC}$	$Ранг^{RD}$	$Ранг^{TC}$	$Ранг^{RD}$	$Ранг^{TC}$	$Ранг^{RD}$
e_1	1	1	2	2	4	3
e_2	4	5	1	1	2	2
e_3	3	3	4	4	1	1
e_4	5	4	5	5	3	4
e_5	2	2	3	3	5	5

Из таблицы 6 очевидно, что если ранжировать узлы в социальной сети и в сети, полученной от первого с помощью расстояния сопротивления применением одной и той же функции, то результаты ранжирования могут различаться. Видно, что в случае выбора коэффициента перекрытия, все участники конференции получили один и тот же ранг, хотя в первом случае при ранжировании используется коэффициент перекрытия, а во втором случае полученный от этого расстояние сопротивления. В отличие от этого, в случае коэффициента Жаккара и нормализованной Google-дистанции картина иная. В первом случае, ранжирование с использованием весов, полученных поисковым агентом, ставит John Miller и Garole Goble в 4-ое и 5-ые места соответственно. Здесь же, при использовании расстояния сопротивления эти два участника обмениваются рангами. Аналогично, в случае нормализованной Google-дистанции Stuart Williams занимает 4-ое и Garole

Goble 3-тьи места, соответственно. Также, при ранжировании с использованием расстояния сопротивления эти два участника меняются местами.

Для тестирования метода выявления сообществ используется следующий эталонный граф:

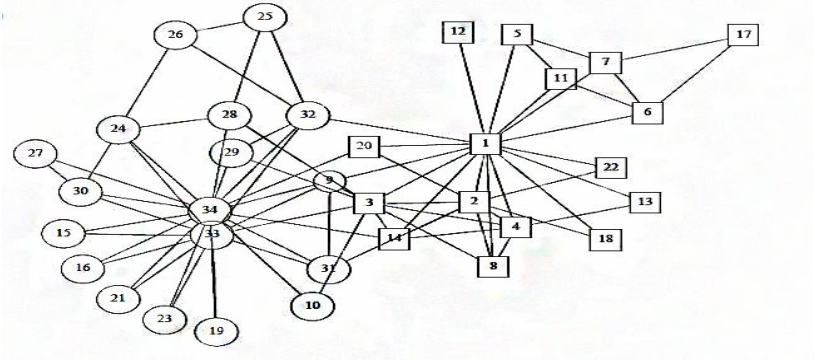

Рис. 2. Социальная сеть клуба каратэ Захари.

Сеть состоит из 34 акторов. Количество кластеров равно 2. В первом кластере имеется 18 акторов, а во втором 16. В результате, с помощью предложенного метода, из 34 акторов 4 (9, 10, 18, 20) были неверно классифицированы (в каждом кластере по 2 актора), а остальные 30 акторов были корректно размещены по кластерам. Такой, результат был получен при значениях параметров $\beta_1=1$ и $\beta_2=0$ в формуле (3). При остальных значениях параметров β_1 и β_2 были получены разные размещения акторов по кластерам. В таблице 7 показаны значения коэффициентов качества кластеризации:

Таблица 7

Значение коэффициентов оценки кластеризации для клуба каратэ Захари

Чистота	Метрика Миркина	F-мера	Коэффициент разбиения	Энтропия	Вариация информации
0,8823	0,2076	0,8824	0,4000	0,3206	0,2053

Как видно из этой таблицы значения коэффициентов дают обнадеживающие результаты. Например, значение чистоты равное 0,8823 очень близко к 1, которая означает, что кластеры, полученные предложенным методом очень близки к реальному. Далее, значение F -меры также близко к 1, что, по определению, показывают высокую точность кластеризации. Кроме того, значение вариации информации указывает на то, что при переходе от действительной кластеризации к предложенной, происходит потеря примерно 20% информации.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В настоящей диссертационной работе получены следующие основные результаты:

1. Разработаны методы синтеза социальных сетей научных исследователей, на основе информации, извлеченной из разных веб-сайтов. Предложенные методы при синтезе результирующей социальной сети учитывают степени важности веб-источников.
2. Разработан метод извлечения социальных сетей научных учреждений, с помощью информации, полученной из нескольких веб-источников.
3. Исследована роль меры близости и алгоритмов ранжирования в анализе социальных сетей, извлеченных на Веб.
4. Предложена оптимизационная модель для выявления сообществ в социальных сетях. Предложен генетический алгоритм для решения оптимизационной задачи.

Основные положения диссертации опубликованы в следующих работах:

1. Ganjaliyev F.S. How similarity measure affects ranking results of network entities / **Azərbaycan Milli Elmlər Akademiyası Aspirantlarının Elmi Konfransı**, Aprel 2010, Bakı, Azərbaycan, 4s.
2. Ganjaliyev F.S. Building a heterogeneous social network of academic researchers / **Proceedings of the 3rd International Conference of Problems of Cybernetics and Informatics**, September 2010, Baku, Azerbaijan, 4p.
3. Alguliev R.M., Aliguliyev R.M., Ganjaliyev F.S. Social network analysis on the web: state of the art, problems and future directions // **İnformasiya Texnologiyaları Problemləri**, 2011, v.3, No1, p.3–11.
4. Ganjaliyev F.S. Building a social network of research institutes on the web / **Azərbaycan Milli Elmlər Akademiyası Aspirantlarının Elmi Konfransı**, May 2011, Bakı, Azərbaycan, 4s.
5. Alguliev R.M., Aliguliyev R.M., Ganjaliyev F.S. Investigation of the role of similarity measure and ranking algorithm in mining social networks // **Journal of Information Science (Impact Factor: 1.299)**, UK, 2011, v.37, No3, p.229–234.
6. Alguliev R.M., Aliguliyev R.M., Ganjaliyev F.S. Extracting a heterogeneous social network of academic researchers on the web based on information

retrieved from multiple sources // **American Journal of Operations Research**, USA, 2011, v.1, No2, p.33–38.
7. Alguliev R.M., Aliguliyev R.M., Ganjaliyev F.S. Aggregating edge weights in social networks on the web extracted from multiple sources with different importance degrees // **Journal of Intelligent Learning Systems and Applications**, USA, 2012, v.4, No2, p.154–158.
8. Alguliev R.M., Aliguliyev R.M., Ganjaliyev F.S. Building a social network of research institutes from information available on the web // **International Journal of Networking and Virtual Organizations**, UK, 2012, v.11, No1, p.62–76.
9. Ganjaliyev F.S. A new method for community detection in social networks extracted from the web / **Proceedings of the 4th International Conference of Problems of Cybernetics and Informatics**, September, September 2012, Baku, Azerbaijan, 2p.
10. Alguliev R.M., Aliguliyev R.M., Ganjaliyev F.S. A partition clustering-based method for detecting community structures in weighted social networks // **International Journal of Information Processing and Management,** Republic of Korea, 2013, v.4, No2, p.60-72.

Роль соискателя в трудах, опубликованных в соавторстве:

[3] – проведен анализ методов и алгоритмов выявления и анализа социальных сетей;

[5] – исследована роль меры близости и алгоритмов ранжирования в социальных сетях;

[6] – предложен метод синтеза социальных сетей научных исследователей;

[7] – предложен метод синтеза социальных сетей научных исследователей;

[8] – предложен метод выявления социальных сетей научных учреждений;

[10] – предложен метод выявления сообществ в социальных сетях.

Fədai Sərəfraz oğlu Gəncəliyev

Veb saytlarda elmi sosial şəbəkələrin aşkarlanması və analizi üçün metod və alqoritmlərin işlənməsi

Xülasə

Dissertasiya işinin məqsədi Veb-dən əldə olunan məlumat əsasında elmi mühitin sosial şəbəkələrinin aşkarlanması və analizi üçün metod və alqoritmlərin işlənməsidir. Bunu nəzərə alaraq dissertasiya işində aşağıdakı nəticələr alınmışdır:

- ✓ Veb-də bir neçə mənbələrdən alınmış məlumat əsasında elmi tədqiqatçıların sosial şəbəkələrinin sintezi üçün metodlar təklif olunmuşdur. Təklif olunan metodlarda məlumat mənbələrin vaciblik dərəcəsini nəzərə almaq mümkündür;
- ✓ Veb-dən alınmış məlumat əsasında elmi tədqiqat qurumların sosial şəbəkələrinin aşkarlanması üçün metod təklif olunmuşdur;
- ✓ Veb-dən aşkarlanmış sosial şəbəkələrdə yaxınlıq ölçüsünün ranqlaşdırma alqoritmlərində rolu araşdırılmışdır;
- ✓ Veb-dən aşkarlanmış sosial şəbəkələrdə icmaların tapılması üçün metod təklif olunmuşdur; metodun həll etmək üçün genetik alqoritm işlənmişdir;

Fadai Sarafraz Ganjaliyev

Development of methods and algorithms for extracting and analyzing academic social networks on the Web

Annotation

The purpose of the dissertation is to develop effective methods and algorithms for extracting and analysis of academic social networks on the Web. According to the specified purpose the following results are received in the dissertation:

- ✓ Methods for merging social network of academic researchers retrieved from different data sources on the Web are developed. Proposed methods for synthesizing of social networks take into account importance degrees of data sources;
- ✓ A method for extracting a social network of research institutes based on information retrieved from multiple sources on the Web is developed;
- ✓ Role of similarity measure in ranking social network actors is investigated and analyzed;
- ✓ A method for detecting communities in social networks extracted from the Web is developed; a modified genetic algorithm is developed for solving the proposed community detection method.

AZƏRBAYCAN MİLLİ ELMLƏR AKADEMİYASI
İNFORMASİYA TEXNOLOGİYALARI İNSTİTUTU

Əlyazması hüququnda

FƏDAİ SƏRƏFRAZ OĞLU GƏNCƏLİYEV

VEB SAYTLARDA ELMİ SOSİAL ŞƏBƏKƏLƏRİN AŞKARLANMASI VƏ ANALİZİ ÜÇÜN METOD VƏ ALQORİTMLƏRİN İŞLƏNMƏSİ

3338.01 – Sistemli analiz, idarəetmə və informasiyanın işlənməsi

Texnika üzrə fəlsəfə doktoru alimlik dərəcəsi
almaq üçün təqdim edilmiş dissertasiyanın

A V T O R E F E R A T I

Bakı – 2013