# Содержательные основания выделения границ Интернет - сетей

**Дарья Вячеславовна Просянюк**

**Национальный исследовательский университет «Высшая школа экономики» (Москва)**

Сегодня Интернет как среда общения, получения и распространения информации, обмена ресурсами и услугами, а также самовыражения стремительно растет, и практика посылает запрос автоматической обработки больших массивов информации. Для поиска информации и принятия решения эксперту, аналитику или простому пользователю уже недостаточно использовать поисковые машины. Для наиболее эффективного решения поставленных задач и поиска необходимой информации нужны новые методы, которые позволили бы пользователю (в широком смысле этого слова) при минимальных усилиях получить необходимую информацию.

Одной из ярких иллюстраций воплощения данной проблемы на практике может быть проблема построения экспертных сетей в Интернете.

Очевидно, что для принятия обоснованных решений необходимо опираться на опыт, знания и интуицию специалистов. В настоящее время все шире применяются различные методы экспертных оценок. Они незаменимы при решении сложных задач оценивания и выбора технических объектов, в том числе специального назначения, при анализе и прогнозировании ситуаций с большим числом значимых факторов - всюду, когда необходимо привлечение знаний, интуиции и опыта многих высококвалифицированных специалистов-экспертов.

Методы экспертных оценок - это методы организации работы со специалистами-экспертами и обработки мнений экспертов, выраженных в количественной и/или качественной форме с целью подготовки информации для принятия решений[1].

Интернет-среда содержит огромное количество хорошо известных, эксплицитно-идентифицируемых сообществ – групп индивидов, объединившихся в формальные группы и разделяющих сходные интересы. Их нахождение и описание не составляет труда. Вместе с тем, существует и не меньшее количество имплицитных сообществ – людей, имеющих сходные интересы, но не объединяющихся вместе, и, возможно, даже незнакомых. Выявление и анализ таких сообществ представляет особый интерес по ряду причин. Сообщества в социальных сетях могут отражать реальные социальные группы; сообщества в сетях цитирования могут представлять связанные статьи на схожие темы; сообщества в метаболических сетях могут представлять циклы или другие функциональные группировки; сообщества во всемирной паутине – страницы на связанные темы. Возможность грамотно идентифицировать

[1] Орлов А.И. Теория принятия решений . Учебное пособие. - М.: Издательство "Март", 2004.

такие объединения поможет понять их свойства и использовать их более эффективно. Следующая причина обусловлена научным интересом. Сообщества репрезентируют социологию сети: их изучение дает представление об интеллектуальной эволюции сети. Другая причина заключается в том, что такого рода сообщества чаще всего концентрируют в себе наиболее ценные и современные информационные ресурсы, необходимые пользователям, интересующимся той или иной тематикой. Четвертая причина состоит в том, что наличие информации о сообществах и их контурах дает возможность для распространения той или иной информации (например, рекламной или идеологической). Наконец, это очень удобный способ работы с экспертным сообществом. Данная проблема особенно актуальна в России, где на сегодняшний день практически невозможно получить независимую информацию от экспертов, не находящихся под влиянием лоббирующих группировок.

Что касается методов выделения границ сетевых сообществ, то традиционным является выделение сообществ в сети путем иерархической кластеризации. Этот метод основан на вычислении силы связей между объектами и поэтапном присоединении (агломеративная кластеризация) или отсоединении (дивизимная кластеризация) объектов. Для вычисления силы связей существует ряд подходов (например, вычисляется общее количество путей между вершинами, или количество независимых путей и пр.).

С другой стороны, существует целый ряд методов, основанных на нечеткой кластеризации объектов. Нечеткая кластеризация зарекомендовала себя как весьма эффективный инструмент анализа данных во многих областях. Например, в биоинформатике она предоставляет возможности для исследования особенностей генной структуры[1].

Наконец, говоря о виртуальных сетях невозможно не упомянуть сравнительно молодой, но динамично развивающийся социолингвистический подход. Одно направление методов данного подхода анализирует синтаксическое содержание файлов, определяя степень их близости[2].

Другое направление центром своего внимания видит не столько содержание файлов, сколько их взаимные связи посредствам гиперссылок[3].

Недостатком существующих подходов является тот факт, что они развивались на стационарных данных. Интернет же является динамической средой, поэтому требует особых подходов и методов анализа. Более того, многие подходы подразумевают, что исследователь изначально осведомлен о границах/совокупности исходных объектов. В реальности же зачастую стоят задачи в определении этих границ.

В связи с этим особое внимание следует уделить двум направлениям исследований – поисково-разведывательному (поиск, описание возможных методов решения задач) и их сравнительному

[1] M. Sato, Y. Sato and L.C. Jain "Fuzzy Clustering Models and Applications," Physica Verlag, Heidelberg, New York 1997.

[2] S. Brin, J. Davis, H. Garcia-Molina. Copy Detection Mechanisms for Digital Documents. Proceedings of the ACM SIGMOD Annual Conference, San Francisco, CA, May 1995.

[3] K. Bharat, Monika R. Henzinger Improved Algorithms for Topic Distillation in a Hyperlinked Environment, 1998.

анализу (подробный анализ методов, сопоставление и подбор к конкретным задачам).

**Поисково-разведывательное направление**

Образование «сгущений» или сообществ (то есть совокупностей узлов с большой плотностью связей друг с другом и с низкой с остальной частью сети) – это естественная характеристика сетевых структур. Конкретные причины образования сообществ могут зависеть от типа сети, но само это свойство является неотъемлемой чертой любой сети, будь то сеть социальная, биологическая или компьютерная. Обнаружение и определение границ таких сообществ является главным шагом на пути к пониманию топологии сети.

Анализ существующих подходов к данной проблематике показал наличие достаточно большого количества методов и способов определения границ сетевых сообществ.

Нами была произведена классификация и подробный разбор основных алгоритмов (см. Таблица 1).

Как видно, все алгоритмы выделения сетевых сообществ могут быть разделены на два большие класса – математические и социолингвистические.

Математические алгоритмы представляет собой широкий класс алгоритмов, разрабатывавшихся не только на виртуальных (шире – текстовых данных), а на более широком круге данных – офф-лайн социальные сети, сети цитирования и соавторства, метаболические и пищевые сети.

Социолингвистические алгоритмы имеют более узкую направленность – будучи разработанными и тестируемыми на текстовых документах, они применимы для анализа текстовой информации (офф-лайн и он-лайн тексты, сайты, гиперссылки и пр.), а также другой информации, теоретически разложимой на совокупности символов: музыка, графика, видео, аудио, базы данных.

**Таблица 1**

**Классификация алгоритмов выделения сетевых сообществ**

| **Алгоритмы выделения сетевых сообществ** | | | | | |
|---|---|---|---|---|---|
| **Математические** | | | | **Социолингвистические** | |
| **Иерархические:** | **Основанные на четкой кластеризации (представители)** | **Основанные на нечеткой кластеризации (представители)** | **Основанные на принципах термодинамики (представи-тели)** | **Анализ синтаксического содержания файлов (представители)** | **Анализ связей файлов (представи-тели)** |
| **Аггломера-тивные** | Besagni D.<br>Boudourides A.<br>Broder A.<br>Polanco X.<br>Garton L.<br>Kumar R.<br>Law J.<br>Rip A.<br>Roche I. | Bezdek J. C.<br>Deva A.<br>Gath I.<br>Gustafson D. E.<br>Jain L. C.<br>Kessel W. C.<br>Palm R.<br>Rusrini E.<br>Sato M.<br>Valdes J. | Zakharov P. | Baker B. S.<br>Brin S.<br>Broader A.<br>Davis J.<br>Garcia_Molina H.<br>Glassman S. C.<br>Heintze N.<br>Manasse M. S.<br>Manber U.<br>Shivakumar N.<br>Zweig G. | Bharat<br>Carrière<br>Henzinger M. R.<br>Kleinberg |
| **Дивизимные** | Djidjev H. N.<br>Girvan M.,<br>Newman M.,<br>Osorio | ... | | | |

Начнем анализ с представителей математического подхода – агломеративных алгоритмов, основанных на четкой кластеризации.

**Применение кластеризации и картографирования веб-сайтов для обнаружения неявных сообществ**[1]

**Задача:** выявление, анализ и визуализация скрытых ассоциаций сайтов.

**Возможности применения:** выявление имплицитных он-лайн сообществ, связанных веб-документов и сайтов.

Статья описывает создание метода анализа ассоциаций веб-сайтов. Этот метод использует связи веб-сайтов для получения представления о структуре ассоциаций. Цель заключается в проведении анализа неявных ассоциаций. Для реализации этого авторы предлагают перейти с совместного анализа слов на совместный анализ сайтов. Сайты считаются связанными, если они оба связаны с третьим сайтом. Авторы предлагают разработанный ими коэффициент значимости ассоциаций, а также картовые и кластерные техники.

Данные, на которых основана эта работа, были собраны в январе 2001 года в Институте компьютерных технологий в г. Патры, Греция, в рамках проекта EICSTES. Первоначально набор данных состоял из 1064 веб-сайтов университетов 22 европейских стран-участниц Европейского Союза. Для каждого научного сайта поисковая система AltaVista определяла количество ссылок с сайта на другие сайты, а также количество внутренних ссылок (то есть ссылок сайта на самого себя). Таким образом была получена квадратная матрица NxN (где N – количество сайтов) и D (i;j) – количество ссылок сайта i на сайт j и D (i;i) – количество внутренних ссылок сайта i. Но в конечном варианте первоначальный набор данных был снижен до 791 университетского веб-сайта из 15 европейских стран.

Алгоритм. Задача web mining, как называют свой подход авторы, состоит из четырех основных этапов: нахождение ресурсов, отбор информации, обобщение и анализ. Два последних этапа описаны в данной работе.

Представленный в статье метод может быть разложен на четыре основные этапа:

1. Трансформация исходной матрицы данных в ассоциативную матрицу;
2. Расчет коэффициента ассоциации;
3. Переформирование сети ассоциаций в кластеры;
4. Размещение полученных кластеров на двумерной карте.
5. Программное обеспечение, используемое в данном исследовании называется SDOC.

[1] X. Polanco, Moses A. Boudourides, Dominique Besagni, Ivana Roche Clustering and Mapping Web Sites. For Displaying Implicit Associations and Visualizing Networks, 2001.

**Масштабируемый многоуровневый алгоритм**
**для кластеризации графов и обнаружения структуры сообществ**[1]

**Задача**: идентификация сетевых сообществ
посредствам кластеризации.

**Возможности применения**: распознавание сообществ в сетевых структурах различной природы: социальные он-лайн
и офф-лайн сети.

Проблема идентификации сетевых сообществ обычно рассматривается с позиции теории графов и кластеров (graph clustering, GC), где вершины представляют отдельные объекты, а ребра описывают отношения между ними. Сообществами тогда считаются подграфы, с наибольшим количеством ребер внутри, и наименьшим – вне (с другими подграфами). Данное исследование представляет новый подход к решению данной проблемы.

Разработанный алгоритм тестировался на нескольких наборах данных. Для сравнения эффективности алгоритма с эффективностью уже известных алгоритмов Ньюмана-Гирвана[2] и Клозета-Ньюмана-Мура[3] использовались случайно сгенерированные графы. Для оценки степени работоспособности и эффективности алгоритма использовались реальные данные домена nd.edu, данные о связях в футбольной команде одного из американских колледжей и в карате клубе Zachary.

**Алгоритм.** Предлагаемый алгоритм разделения графа состоит из следующих фаз:

1 *фаза – грубое разделение – coarsening phase.*
Исходный граф G разделяется на подграфы и каждый из них заменяется одной вершиной, а множество ребер, связывающих эти подграфы – одним ребром. Вес каждой новой вершины (и, соответственно, ребра) равен сумме весов вершин (и, соответственно, ребер), которые они заменили. Таким образом деление графа повторяется несколько раз до тех пор, пока его размер не станет сравнительно малым. Пусть G0 = G, G1, …, Gl – итоговая последовательность графов.

2 *фаза – разделение – partitioning phase.*
Граф Gl разделяется на две части при помощи любого доступного метода (например, спектральное разделение по алгоритму Кернигана – Лина[4].

3 *фаза – деликатное разделение и очищение – uncoarsening and refinement phase.*
Данная фаза является наиболее важной в алгоритме, так она в большей степени именно она определяет уровень его точности и эффективности, поэтому она будет описана несколько более подробно. Планируемое разделение Gl равно Gl-1. Так как вес каждой вершины Gl равен сумме весов соответствующих вершин Gl-1 , то часть Gl-1 будет уравновешена, если часть Gl и отрезы обоих частей будут иметь одинаковые веса.

[1] H. N. Djidjev A scalable multilevel algorithm for graph clustering and community structure detection, Los Alamos National Laboratory, Los Alamos, NM 87545.
[2] M. Newman and M. Girvan. Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004.
3 A. Clauset, M. Newman and C. Moore. Finding community structure in very largenetworks,
Phys. Rev. E 70, 066111 (2004).
4 Kerninghan B. W. and Lin S. An efficient heuristic procedure for partitioninggraphs, The Bell System Technical Journal, 1970).

Однако, поскольку в Gl-1 входило большее количество вершин, чем в Gl, она имела большее количество степеней свободы и, следовательно, возможно отделить часть Gl-1 с тем, чтобы снизить её размер (длину кратчайшего пути).

Алгоритм Кернигана – Лина описывается набором итераций, каждая из которых состоит в перемещении вершин из одного подграфа в другой.

Реализация данного алгоритма возможна в программном пакете METIS.

Для сравнения эффективности данного алгоритма и алгоритма Ньюмана-Гирвана был сгенерирован произвольный граф с 128 вершинами и 4 сообществами из 32 вершин каждое. Предполагаемая степень каждой вершины равна 16, но внешняя степень (то есть ожидаемое количество смежных вершин, принадлежащих к другому сообществу) изменяется от 1 до 8. Следовательно, наибольшие значения внешних степеней присущи вершинам, принадлежащим слабосвязанным кластерам.

Данный алгоритм оказался эффективнее алгоритма Ньюмана-Гирвана при любых значениях внешних степеней вершин.

При сравнении с алгоритмом Клозета-Ньюмана-Мура (который отличается от алгоритма Ньюмана-Гирвана более короткими сроками реализации), метод также показал превосходные результаты.

Также алгоритм был протестирован на нескольких наборах реальных данных. Во всех случаях алгоритм продемонстрировал практически безошибочные результаты распознавания сообществ.

Итак, исследование представляет собой разработку нового подхода к кластеризации графов. Алгоритм показал свою эффективность как при тестировании на данных с уже известной структурой, так и при сравнении с уже известными алгоритмами.

**Недостатки:** Основным недостатком этого подхода является отсутсвие описания реализации подстадий (фаза 1 – грубое разделение – «Исходный граф G разделяется на подграфы …») Как? Очевидно, что для решения данной задачи может быть предложено несколько вариантов (которые, вероятнее всего, способны составить отдельную стадию исследования). Очевидно, что от выбора конкретного способа зависят результаты, а, следовательно, и дальнейшая реализация алгоритма. Авторам следовало бы указать возможные способы разделения графа на подграфы, или, как минимум, способ, использованный ими.

**«Быстрый» алгоритм Ньюмана**
**для распознавания структуры сообществ в сетях**[1]

**Задача:** выявление кластеров в сетевой структуре.

**Возможности применения**: распознавание сообществ в сетевых структурах различной природы: социальные он-лайн и офф-лайн сети.

Авторы предлагают алгоритм, основанный на итеративном отсоединении элементов с наибольшей центральностью. Данный подход был опробован на различных видах сетей, таких как e-mail сообщения, социальные сети животных, объединения джазовых музыкантов, генные разработки и пр.

1 M. E. J. Newman Fast algorithm for detecting community structure in networks, 2003.

**Алгоритм.** Алгоритм основан на идее «модулярности». В любой сети алгоритм формирует некоторые объединения вершин, вне зависимости от того существуют ли они на самом деле. Для определения качества решения используется специальная функция – «модулярность», рассчитываемая по формуле:

$Q = \Sigma i\ (eii - a^2 i)$, где

eii – это все ребра, соединяющие каждую вершину кластера i с другими вершинами этого же кластера,

ai – это все ребра, выходящие из каждой вершины кластера i и связывающие её с вершинами кластера j.

Таким образом, в ситуации случайного объединения вершин Q=0 и в этом случае есть основания говорить об отсутствии выделенных сообществ. Величина Q изменяется от 0 до 1 и на практике её значения, превышающие 0,3 свидетельствуют о наличии сообществ.

Однако проблема максимизации значения Q обычно является слишком затратной для решения и на практике пригодна лишь для случая наличия 20-30 вершин. Для решения этой проблемы авторы предлагают следующий оптимизирующий алгоритм. По свой принадлежности он попадает в категорию агломеративных иерархических кластерных методов. На первом этапе каждая вершина является отдельным сообществом. На следующем этапе вершины объединяются в пары. Цель – наибольшее увеличение (или наименьшее снижение) значения Q. Процесс реализации алгоритма может быть визуализирован в помощью дендрограммы. «Срезая» дендрограмму на разных уровнях мы имеем возможность варьировать количество сообществ.

Очевидно, что для увеличения значения Q необходимо рассматривать только те сообщества (вершины), между которыми существуют ребра. Изменение функции Q рассчитывается по следующей формуле:

$\Delta Q = eij - eji - 2aiaj = 2(eij - aiaj)$, где

eij – половина ребер, соединяющих кластеры i и j,

eji – другая половина ребер, соединяющая кластеры j и i.

**Недостатки:** одним из значимых недостатков алгоритма авторы называют значительные временные затраты для его реализации.

**Структуры сообществ в биологических и социальных сетях**[1]

**Задача:** выявление границ сообществ в сетевых структурах.

**Возможности применения:** идентификация сообществ в сетевых структурах различной природы: социальные сети, сети цитирования и соавторства, пищевые сети животных, нейронные сети, компьютерные и он-лайн сети.

Традиционным методом распознавания сообществ в сети является иерархическая кластеризация – пошаговое объединение наиболее сильно связанных узлов. Для расчета силы связей предлагаются различные методики. Однако такой способ имеет ряд недостатков – игнорирование наиболее периферийных элементов. Также, при тестировании этого метода на сетях с уже известной структурой и границами сообществ он показал весьма скромные результаты. Поэтому авторами предлагается альтернативный подход. Вместо определения наиболее близких узлов,

1 M. Girvan, M. E. J. Newman. Community structure in social and biological networks, 2002.

метод предлагает выявлять наименее близкие элементы сети. То есть вместо того, чтобы конструировать сообщества, присоединяя узлы, метод конструирует их, пошагово отсоединяя от исходного графа.

В исследовании используются 2 типа данных: сетевые структуры с уже известными границами сообществ и сети, в которых сообщества еще не были распознаны. В первую группу входили искусственно сгенерированное множество графов, данные о социальных связях членов каркте клуба, а также социальная сеть американской футбольной команды. Во вторую группу входили данные о соавторстве членов исследовательского центра в Санта Фе (Нью Мексико).

**Алгоритм.** Предложенный алгоритм выглядит следующим образом:

1. Расчет посредничества центральности (betweenness) для всех ребер сети (Авторы вводят понятие посредничества центральности ребра - количество кратчайших путей между парами вершин, которые проходят параллельно (рядом, в одном направлении) этому ребру.);
2. Удаление ребер с наибольшим значением посредничества центральности;
3. Пересчет посредничества центральности для ребер, связанных с удаленными;
4. Возвращаемся ко второму шагу и продолжаем до тех пор, пока не останется ни одного ребра.

Для расчета посредничества центральности ребер авторы предлагают использовать быстрый алгоритм Ньюмана, описанный выше.

Итак, вышеописанный метод выделения сетевых сообществ опирается на алгоритм Ньюмана и предлагает изначально определять периферию сети. При тестировании на сети с изначально известной структурой метод показал блестящие результаты, практически безошибочно её распознав. В случае с неизвестной топологией сети результаты метода подтвердили изначальные предположения и послужили толчком для дальнейших изысканий. Так, например, при исследовании научного сообщества были выделены 2 сообщества ученых: по сходной исследовательской тематике и по сходной методологии исследований.

В дальнейшем авторы предлагают усовершенствовать метод, введя в анализ веса и направленные графы. Также возможно применить разработанный метод в исследовании других типов сетей, например, нейронных или виртуальных.

**Недостатки:** Недостатком данного алгоритма являются (по мнению самих авторов) значительные временные затраты на его реализацию.

### Интерпретация результатов нечеткой кластеризации с помощью их виртуального представления с использованием генных микрочипов[1]

**Задача:** нечеткая кластеризация объектов, визуализация и анализ результатов кластеризации.

**Возможности применения:** нечеткая кластеризация и анализ структуры комплексных, неполных и неточных данных: медицинских, биологических, он-лайн.

1 Valdes, J. Interpreting Fuzzy Clustering Results With Virtual Reality-based Visual Data Mining: Application to Microarray Gene Expression Data. March 2004.

Как показывают многие исследования, нечеткая кластеризация является весьма эффективным инструментом анализа данных. Данная работа предлагает алгоритм нечеткой кластеризации разнородных структур и демонстрацию его реализации на виртуальных данных.

Виртуальная реальность представляет достаточно простые, но в то же время мощные возможности для понимания и интерпретации комплексных, неполных и неточных данных. Подход, описанный в данной статье, апробировался на базах медицинских данных о генах болезни Альцгеймера и лейкемии.

В изучении болезни Альцгеймера было взято 23 пробы больных и здоровых людей. Эти пробы были описаны 9600 генами. В результате работы алгоритма были получены 2 четких и 2 нечетких кластера, m = 4, коэффициент разделения равен 0,571556, множество геометрий визуального пространства G = {сфера, конус, куб}.

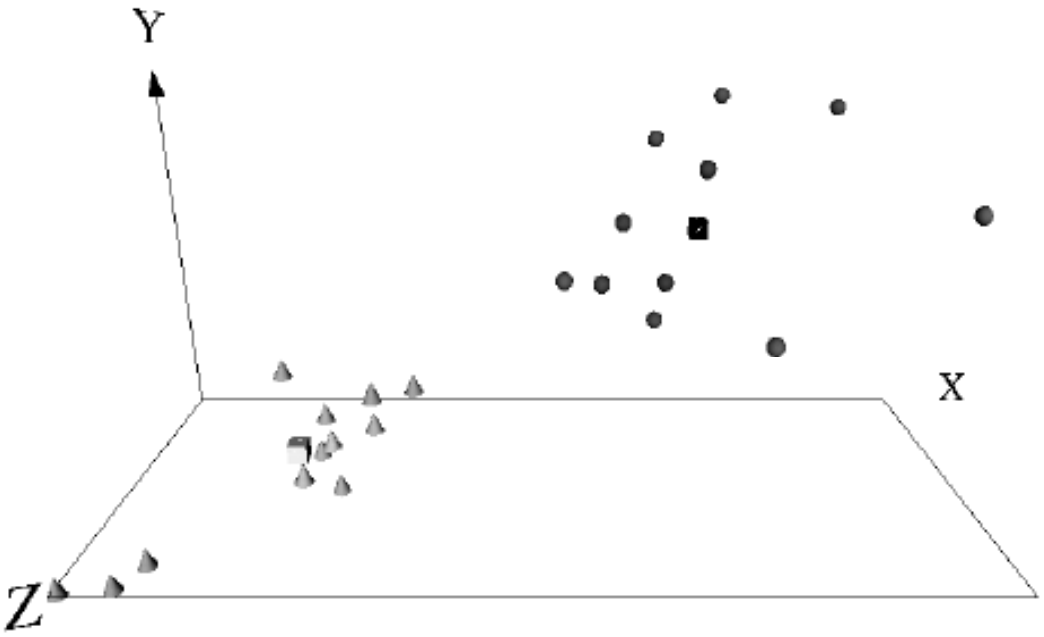


**Рисунок. Визуальное представление получившихся кластеров.**

Сферы и конусы представляют четкие кластеры (гены болезни Альцгеймера (белый цвет) и нет (черный цвет), кубы представляют расположение центров нечетких кластеров (серый цвет).

Подобные результаты были получены и в случае исследования лейкемии.

**Алгоритм.** Цель нечеткой кластеризации – формирование кластеров объектов, основанное на их сходстве, которое определяется по их атрибутам, а также по заданному формальному критерию оценки сходства (различия). При нечетком разделении n объектов на K кластеров кластерная структура может быть описана матрицей U = (uik) размера n x K, где uik принадлежит [0,1], i = 1, …, n; k = 1,…, K и необходимо, чтобы Σ k=1K uik = 1. uik представляет принадлежность каждого объекта к кластеру. Значение близкое к 1 свидетельствует о сильном сходстве между объектами, близкое к 0 – о слабом сходстве. Такой подход демонстрирует классические основы нечеткой кластеризации, так как один объект может всецело принадлежать одному кластеру или же –

частично нескольким. Мерой оценки качества кластерного решения служит сумма внутрикластерных дисперсий. Достижение хорошего кластерного решения подразумевает ее минимизацию.

Классический алгоритм продолжается последующей аппроксимацией изначально оцененных центров:

Затем членство каждого объекта определяется по специальной формуле.

Проблема оптимизации Jm довольно сложна для решения. Обычно, для гарантии проверки кластерного решения в дополнение к Jm используются дополнительный коэффициенты – коэффициент разделения Fc и коэффициент энтропии Hc.

Итак, данный подход позволяет отображать в виртуальной реальности одновременно четкие и нечеткие кластеры. Возможности, предоставляемые данным подходом, практически бескрайние, поскольку существует неисчислимое множество различных мер близостей, разностей и расстояний.

Данный подход является единственным представителем подкласса математических алгоритмов и предполагает использование законов физики (явление диффузии, закон сохранения массы) для решения социологических задач - распознавания сообществ в сети.

**Термодинамический подход для распознавания сообществ в комплексных сетях: Исследование Живого Журнала**[1]

**Задача:** выявление (очерчивание границ) крупных сетевых сообществ

**Возможности применения:** выявление сообществ в социальных (в частности, он-лайн) социальных сетях.

Как известно, одной из основных характеристик Живого Журнала является возможность «дружбы» пользователей. Каждый пользователь Живого Журнала имеет свой список друзей, куда он добавляет других пользователей (исходящие связи), а также он может быть добавлен в список друзей другого пользователя (входящие связи). Если пользователь А присутствует в списке друзей пользователя Б, и имеет пользователя Б в списке своих друзей, то такая связь является взаимной. Для данного исследования собирались данные о пользователях Живого Журнала, включающие их Ники (виртуальные имена), входящие, исходящие и взаимные связи. Время сбора данных составило 14 дней. Количество исследуемых пользователей – 3746264. Среднее количество исходящих связей Кисх = 15,91, входящих Квх = 16,07, среднее соотношение входящих и исходящих связей (Квх/Кисх = 1,157). Средний коэффициент кластеризации (который представляет собой отношение связей друзей пользователя к максимально возможному количеству связей между ними) на данных равен 0,3302, что свидетельствует о высокой степени пользовательской кластеризации (для сравнения, коэффициент кластеризации случайного направленного графа с теми же параметрами, что и исследуемый равен 4, 24 x $10^{-6}$.

1 Zakharov P. Thermodynamic approach for community discovering within the complex networks: LiveJournal Study, 2008.

**Алгоритм.** Метод обнаружения сетевых сообществ, предлагаемый в данном исследовании, основан на принципах термодинамики. Основная идея – имитация процесса диффузии в комплексных сетях как в многомерных «пористых» системах с направленными связями, подчиняющихся физическим законам. Процесс диффузии инициируется в одном из узлов путем добавления в него «виртуальных чернил», распространяющихся затем по другим узлам. Сильно связанные между собой узлы (другими словами – сообщества) идентифицируются по одинаковому количеству чернил в них. В этом смысле данный метод может быть отнесен к категории агломеративных методов.

Итак, прежде всего, предположим, что узел имеет неограниченную емкость. Прямой поток из узла А в узел Б возможен только если между А и Б есть направленная связь и фА > фВ (где фА и фВ – это значения концентраций «чернил» в узлах А и Б). Сетевые связи в этом случае представляются трубками, направленные связи – трубки, распространяющие «чернила» только в одном направлении.

Теперь представим себе кластер узлов, плотно связанных друг с другом, но имеющий всего несколько внешних связей с остальной частью сети. Тогда диффузия чернил будет происходить сравнительно быстро внутри этого кластера. Небольшое же количество внешних связей ведет к эффекту «бутылочного горлышка» - сравнительно медленной скорости обмена чернилами узлов кластера с внешними узлами. Таким образом, члены одного кластера будут иметь одинаковое количество «чернил», измерение которого позволит идентифицировать эти кластеры.

Данный метод не предназначен для разбиения сети на маленькие совокупности, он помогает распознать сравнительно большие и значимые кластеры. Используя логику и терминологию Ньюмана, авторы предлагают для оценки качества решения следующий коэффициент:

$$K_i = \frac{e_{ii}}{\sum_j e_{ij}} = \frac{e_{ii}}{b_i}$$

, где

eij – доля ребер сети, связывающих кластеры i и j,

$\sum_j e_{ij} = b_i$ - доля ребер сети, имеющих начало в узлах сообщества i.

Таким образом, этот параметр оценивает долю внутренних связей сообщества i среди всех связей, исходящих из сообщества i.

Для тестирования метода было решено использовать несколько разных начальных узлов (юзеры doctor_livsy – писатель Сергей Лукьяненко и future_visions). В результате было обнаружено 2 большие сообщества, сравнительно слабо связанные между собой – русскоговорящий сегмент Живого Журнала и остальная часть сети.

**Недостатки:** Что такое «виртуальные чернила»? Из статьи этого не ясно. Сам автор так отвечает на этот вопрос: «виртуальные "чернила" - это просто абстрактная субстанция, необходимая для моделирования диффузии в подобной сети. Однако, информацией в моей конкретной модели она быть не может, поскольку для нее не действует закон сохранения массы. В моем блоге я давал объяснение принципа работы алгоритма: "Объясняю на пальцах, то есть на деньгах, что есть эта самая термодинамичесая дистанция: дал я [info]doctor_livsy некую сумму денег

в рублях и сказал: распределяй каждый день (допустим) деньги между своими друзьями так, чтобы у вас у всех было поровну. Потом пусть друзья тоже присоединяются и раздают своим друзьям, а дальше - их друзья и т.д. То есть каждый день все у кого есть деньги делятся с теми друзьями, кто в них нуждается. В определенный момент я говорю стоп и проверяю сколько у кого денег. Так вот, разница в накоплениях и определяет расстояние на карте. Допустим теперь [info]future_visions одновременно раздает другую валюту - доллары, например. Тогда их сумма определяет вторую координату. Посмотрите теперь, что получается - поскольку российский сектор хорошо связан внутри, но слабо соединен с остальным LJ лишь малая часть рублевой массы уходит вовне, в то время как среди РЖЖ юзеров рубли распределены почти равномерно, а с долларами получается обратно - они слабо приходят в российский сектор, но быстро распределяются внутри него. Из-за этого и получается отдельный остров».

Следующим недостатком является слишком высокая степень вовлеченности исследователя: для определения границ сети ему самому необходимо предпринимать определенные весьма затратные действия: распространять «чернила». Более того, непонятно по какому принципу определять инициатора (начальный узел добавления «чернил»), и что зависит от этого инициатора?

Следующие статьи представляют социолингвистическое направление анализа.

**Обнаружение сходных файлов**
**в крупной файловой системе**[1]

**Задача:** распознавание текстовых файлов из одного ресурса или содержащих одинаковых фрагменты текста.

**Возможности применения:** Основная цель метода – распознавание текстовых файлов из одного ресурса или содержащих одинаковые фрагменты текста. Два файла считаются сходными, если они содержат значительное количество общих цепочек знаков. Файлы не обязательно должны быть одинакового размера, один может включать другой.

Метод предлагает новый подход, основанный на понятии так называемых «примерных отпечатков». «Примерный отпечаток» представляет собой компактную репрезентацию файла такую, что «отпечатки» двух схожих файлов будут схожими (не обязательно одинаковыми), а «отпечатки» двух разных файлов будут разными. Алгоритм (исходное название которого sif – Прим. авт.) может быть реализован двумя способами: все-против-всех и один-против-всех). В первом случае алгоритм выбирает все сходные файлы из заданного множества и выявляет степень их схожести. Второй способ сравнивает один заданный файл с предварительно сформированным каталогом всех других файлов и очень быстро (напр., 3 сек. для 4000 файлов размером 60 Mb) определяет те файлы, которые похожи на заданный. В обоих случаях схожесть может быть установлена, даже если сходный фрагмент составляет всего 25% размера меньшего файла.

1 U. Manber. Finding similar files in a large file system. Proceedings of the 1994 USENIX Conference, pp. 1-10, January 1994

Авторы предсказывают несколько возможных применений разработанного ими алгоритма sif. Это значительно облегчает выполнение задач так называемому файл-менеджменту (обнаружение разных версий одних и тех же программ и статей), помощь в системном администрировании организаций, разработка поисковых механизмов в Интернете, сжатие сходных файлов. Также данный метод может быть полезен простым пользователям, которые, например, используют несколько компьютеров (положим, домашний, рабочий и портативный). Преподавателям и писателям он может помочь выявить плагиат при написании работ, политическим деятелям – сходные фрагменты официальных писем и деловой корреспонденции и пр.

Альтернативный подход выявления схожести файлов был предложен Б. Бейкер. Она предлагала считать два файла схожими, если один может быть получен из другого изменением параметров слов. Автор называла такой способ проверки на схожесть параметрическим согласованием и предлагала несколько алгоритмов для выявления схожести файлов.

**Алгоритм.** Для выявления степени схожести файлов предлагается формировать «отпечатки» файлов, причем несколько для каждого (однако не стоит в качестве оных использовать части текста).

Для создания «отпечатков» текстов предлагается 2 способа. Первый из них – якорный. Якорем называется последовательность символов. При начале каждой процедуры сравнения файлов используется постоянное множество якорей. Если два файла содержат идентичный фрагмент, и этот фрагмент содержит якорь, тогда последовательность символов около якоря также идентична. Например, предположим, что якорем мы выбрали последовательность символов арак. Находим все файлы, содержащие последовательность символов арак. Например, это файлы, содержащие слово характеристика. Затем мы внимательно изучаем каждый текст на предмет содержания в нем последовательности символов, включающих арак. Выбираем фиксированное число этих последовательностей, скажем, 50. Эта контрольная сумма и будет «отпечатком» файла. Такие «отпечатки» будут сформированы для всех файлов, содержащих эти 50 единиц последовательностей. В случае, если в файле нет 50-ти последовательностей символов, содержащих арак, «отпечаток» для него не будет сформирован. Тонкость такого подхода состоит в использовании нескольких якорей для описания файлов однообразным способом.

Второй способ создания «отпечатков» текстов представляется авторов более простым и удобным. Его идея состоит в формировании «отпечатков» из всех возможных подстрок определенной длины и дальнейшем выборе подмножества этих «отпечатков» на основании их весов.

Задача выбора «отпечатка» имеет несколько решений. Простейшее из них состоит в выборе того, чьи последние k бит равны 0. Приблизительно 1 «отпечаток» будет выбран из 2k последовательностей.

*Сравнение один-против-всех.* Этот способ, как уже было обозначено выше, сравнивает один файл (назовем его запрашиваемый файл) с множеством файлов, для которых заблаговременно уже были сформированы «отпечатки». Набор всех «отпечатков» составляет отдельный файл. Для запрашиваемого файла также составляются

«отпечатки» (также отдельным файлом). Следующий шаг заключается в сравнении данных двух файлов, содержащих «отпечатки».

*Сравнение все-против-всех.* Этот способ представляется более сложным. На первом этапе для всех файлов (как и в предыдущем способе) составляются «отпечатки». Затем для каждого «отпечатка» составляется список файлов, содержащих его. Далее представление информации немного видоизменяется – для каждого из выбранных множеств файлов указывается количество общих «отпечатков». Пользователю необходимо задать порог количества общих «отпечатков» по достижении которого файлы будут считаться сходными.

*Возможные направление будущих разработок.*

Области применения представленного метода могут быть расширены как минимум в четырех направлениях:

- Расширение возможных типов файлов – комментарии (пояснения, сноски и пр.), рабочие и сжатые файлы;
- Дробление и исправление файлов различных размеров;
- Усовершенствование механизмов сравнения двух файлов;
- Улучшение характеристик выходных данных.

**Недостатки:** Алгоритм не работает с содержательным составом файлов. Он сосредоточен на синтаксическом составе. То есть файлы, содержащие сходную информацию, но написанные разными словами не считаются сходными. Другими словами данный подход не может использоваться для выявления семантической близости текстов.

**Синтаксическая кластеризация всемирной паутины**[1].

**Задача:** анализ схожести символического состава файлов, степени вмещения одного файла другим. Объединение схожих файлов в кластеры.

**Возможности применения:** Алгоритм разрабатывался и тестировался на текстовых документах. Также он может применяться для анализа:

- Аудио-сообщений, содержащих человеческую речь;
- Документов на иностранных языках;
- Музыкальных композиций;
- Картинок, видео, баз данных.

Более того, этот метод может использоваться для распознавания измененных URL-страниц, кластеризации результатов поиска в поисковых системах для более удобного их использования, анализа изменения сайтов, а также для распознавания плагиата.

Предложенный метод позволяет определять синтаксическую схожесть файлов в Интернете. Разработанный алгоритм основан на установлении степени совпадения символического состава файлов, а также степени вмещения одного файла другим.

Исходные данные исследования содержали 30000000 HTML и текстовых документов из сети Интернет. Парное сравнение содержало $10^{15}$ (квадриллион) сравнений. Окончательная кластеризация происходила на основе 50%-й степени схожести. Было обнаружено 3,6 миллиона кластеров, содержащих 12,3 миллионов документов (из них 2,1

1 Andrei Z. Broader, Steven C. Glassman, Mark S. Manasse, Geoffrey Zweig. Syntactic Clustering of the web. 1997

миллион кластеров содержали идентичные документы (5,3 миллиона документов), 1,5 миллиона кластеров содержали 7 миллионов документов).

**Алгоритм.** Каждый документ рассматривается в первую очередь как последовательность слов. Назовем смежную последовательность символов во множестве всех символов документа D *«шинглом»* (shingle). Каждому документу D ставится в соответствие множество последовательностей символов S (D, w). То есть каждый исходный документ D мы определяем как w-шингловый как множество всех уникальных шинглов размера w. Так, например, 4-шингловое множество для документа {эта, роза, красива, эта, роза, красива, эта, роза} будет множество {эта, роза, красива, эта}, {роза, красива, эта, роза} и {красива, эта, роза, красива}.

Для определения степени схожести документов A и B использовались две метрики: *сходство* (resemblance) и *вместимость* (containment).

Для данного шингл-размера сходство рассчитывается по следующей формуле:

$$r(A; B) = |S(A) \cap S(B)| / |S(A) \cup S(B)|$$

Показатель сходства документов изменяется в интервале [0;1] – чем ближе эта величина к единице – тем более документы похожи.

Вместимость документов рассчитывается как

$$c(A; B) = |S(A) \cap S(B)| / |S(A)|$$

Подобным образом измеряется и вместимость – от 0 до 1. Чем ближе она к единице, тем более один документ включен в другой.

Для расчета данных показателей удобно иметь *эскиз* (sketch) (то есть некоторую «выжимку») каждого документа на несколько сотен байт. Имея эскизы двух документов по ним становится возможным рассчитать вышеописанные метрики.

Пусть U – это множество всех «шинглов» зафиксированного размера w. Параметр s определим следующим образом: для множества W C= U определим MINs (W) как

$$\begin{cases} \text{MINs (W) = множество минимальных s элементов в W, если } |W| >= s; \\ W \quad \text{, иначе.} \end{cases}$$

где «минимальный» относится к цифровому порядку и определяется MODm(W) = множество элементов W, от 0 до W.

Пусть $\pi$ : U $\longrightarrow$ U это случайное равномерное изменение U. Пусть F(A) = MINs($\pi$(S(A))) и V(A)=MODm($\pi$(S(A))). F(B) и F(A) определим аналогично.

Тогда

Полезность = $|MINs(F(A) \cup F(B)) \cap F(A) \cap F(B)| / |MINs(F(A) \cup F(B))|$ - и

Полезность = $|V(A) \cap V(B)| / |V(A) \cup V(B)|$

являются объективными оценками схожести документов A и B.

Полезность = $|V(A) \cap V(B)| / |V(A)|$ - объективная оценка вместимости документов A и B.

Алгоритм имел следующий вид:

1) Извлечь информацию из каждого документа;
2) Сформировать эскиз каждого документа;
3) Сравнить эскизы каждой пары документов на предмет превышения порога схожести;
4) Сформировать пары схожих документов с целью их дальнейшего объединения в кластеры;

Кластеризующий алгоритм имеет следующий вид:

1) На первом этапе формируем эскиз каждого документа;
2) На втором этапе создаем список всех «шинглов» и документов, в которых они присутствуют, сортируя их по значимости. Вследствие этого эскиз каждого документа расширяется на пару параметров <шингл, ID документа>;
3) На третьем этапе мы формируем список всех пар документов, содержащих идентичные «шинглы», указывая какое количество общих «шинглов» они имеют. Таким образом, список <шингл, ID документа> расширяется до <ID, ID, количество общих шинглов>.
4) На финальной стадии формируем кластеры схожих документов. Анализирую триады <ID, ID, количество общих шинглов> мы принимаем решения о том, превышает ли степень схожести документов установленный нами порог. Если да – устанавливаем связь между двумя документами и относим их к одному кластеру.

Итак, статья содержит описание метода, позволяющего выделять схожие/идентичные множества элементов (символов) и измерять степень их схожести, объединяя затем в кластеры.

Сходная проблематика описана в статье Nevin Heintze. *Scalable Document Fingerprinting*[1], где акцент делался на распознавании плагиата. Также данной проблематикой занимались U. Manber[2], N. Shivakumar и H. Garcia-Molina.[3], S. Brin, J. Davis[4].

Следующий алгоритм своей основой имеет связи (гиперссылки) между документами.

[1] Nevin Heintze. *Scalable Document Fingerprinting.* Proceedings of the Second USENIX Workshop on Electronic Commerce, Oakland, California, November 18-21, 1996

2 U. Manber. Finding similar files in a large file system. Proceedings of the 1994 USENIX Conference, pp. 1-10, January 1994.

[3] N. Shivakumar, H. Garcia-Molina. *SCAM: A Copy Detection Mechanism for Digital Documents.* Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries, Austin, Texas, 1995

[4] S. Brin, J. Davis, H. Garcia-Molina.*Copy Detection Mechanisms for Digital Documents.* Proceedings of the ACM SIGMOD Annual Conference, San Francisco, CA, May 1995.

**Улучшенный алгоритм**
**«отсеивания» тем в он-лайн среде**[1]

**Задача:** поиск документов наиболее релевантных теме запроса.

**Возможности применения:** выявление публикаций и авторов со сходными интересами, заказных публикаций, плагиата

Статья описывает проблему «отсеивания» тем в Интернет – поиск документов, наиболее релевантных запросам пользователей. Подход представляет собой объединение анализа связей документов и контент-анализа. Авторы описывают три основные проблемы в данном поле - взаимное усиление позиций связанных документов, автоматическое генерирование ссылок и выдача нерелевантных документов – и предлагают методы их решения.

**Алгоритм. Анализ связей.** Основная цель анализа связей документов – использование гиперссылки между документами для отбора необходимых документов (основываясь на предположениях, что документы, связанные гиперссылками имеют сходное содержание и что если один автор в своей работе дает ссылку на работу другого, то он считает эту работу ценной). Рассмотрим подробнее алгоритм Клейнберга, представляющий данный метод[2].

*Улучшенный алгоритм Клейнберга.*

Введем две метрики документа: *центральность* и *авторитетность*. Документы с большим значением авторитетности хорошо соответствуют теме запроса, в то время как документы с большим значением центральности содержит большое количество ссылок на документы, релевантных теме. Далее вытекает следующая логика: документ, который ссылается на многие документы имеет высокую центральность, документ, на который ссылаются многие документы – высокую авторитетность. Транзитивно, документ, который ссылается на многие авторитетные документы еще более централен, так же, как и еще более авторитетен тот, на который ссылаются многие центральные. В контексте запроса пользователя алгоритм вначале формирует определенный граф, где вершинами являются документы. Затем узлам итеративно назначаются центральности и авторитетности. Граф формируется следующим образом: стартовое множество документов, релевантных запросу, определяется поисковиком (скажем, 200 первых выданных документов). Затем это множество увеличивается на «соседей» этих документов – то есть документы которые ссылаются на и на которые ссылаются выданные документы (автор советует рассматривать не более 50 предшественников каждого документа). Эти два множества формируют соседский граф. Ребра графа - гиперссылки между документами.

Если документы одного источника связаны k связями с одним документом второго источника, то каждой связи присваивается вес авторитетности (authority weight) 1/k. Если один документ из первого источника связан l связями с документами второго источника, то для каждой связи вес центральности (hub weight) – 1/l. Изолированные узлы не рассматриваются.

Расчет центральностей и авторитетностей документов происходит по следующей схеме:

1) Пусть N – это множество вершин соседского графа;

1 K. Bharat, Monika R. Henzinger Improved Algorithms for Topic Distillation in a Hyperlinked Environment, 1998.

2 Kleinberg, J. 1998. \Authoritative sources in a hyperlinked environment." Proc. of 9th ACMSIAM Symposium on Discrete Algorithms. Also appeared as IBMResearch Report RJ 10076, May 1997.

2) Для каждой вершины n из множества N, H[n] – ее центральность, A[n] – ее авторитетность;
3) Назначим H[n] и A[n] равными 1 для всех n из множества N;
4) До тех пор пока H[n] и A[n] не сойдутся в одну точку:
5) Для всех n из множества N, A[n] = $\sum_{(n',n)\in N} H[n'] \times auth\_wt(n',n)$;
6) Для всех n из множества N, H[n] = $\sum_{(n,n')\in N} A[n'] \times hub\_wt(n,n')$;
7) Нормируем вектора H[n] и A[n].

По уверениям авторов, алгоритм сходится после примерно 10 итераций.

Далее, для определения релевантности документов, авторы предлагают совместно использовать анализ связей и контент-анализ.

1) Очевидно, что тема запроса шире, чем сам запрос. Поэтому авторы предлагают использовать документы стартового множества как расширенный запрос и сопоставлять каждый узел (документ) графа с ним (точнее, с первыми 1000 словами каждого документа).
2) Для удаления нерелевантных документов предлагается три методики:
   1. Медиана. Порог – медиана всех значимых весов;
   2. Медиана стартового множества. Порог – медиана значимых весов узлов стартового множества;
   3. Доля максимальных весов. Порог – фиксированная доля максимальных весов. В данном исследовании использовался порог макс/10.

Алгоритм ARC Чаркабарти при анализе текстов также опирается на алгоритм Клейнберга. Но ARC-алгоритм, в отличие от данного, использует соседство второй очереди.

Анализ связей используется при исследовании соавторства в библиометрике.

**Финальное сравнение алгоритмов**

В таблице 2 представлены результаты финального сравнения алгоритмов. Как видно, нами был выделен ряд критериев для сравнения. Отдельной частью являются недостатки алгоритмов, которые по своей природе не являются параметрами для сравнения, а представляют собой характеристики алгоритмов.

**Таблица 2**

**Финальная сравнительная таблица**

| **Решаемая задача** | **Выявление степени схожести текстов (схожих, вмещенных файлов)** | | **Поиск документов, релевантных теме запроса** | **Выявление сообществ** | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | **Четкие границы, любой размер** | | | **Нечеткие границы, крупные размеры** | |
| **Алгоритм (фамилия автора)** | **Broader** | **Manber** | **Bharat** | **Girvan, Newman** | **Djidjev** | **Polanco** | **Zakharov** | **Valdes** |
| Декларируемые области применения | Анализ текстов, аудио, видео, графики | Файл-менеджмент, поиск, преобразование данных | Выявление авторов со сходными интересами, заказных публикаций, плагиата | Сети: социальные, сети цитирования и соавторства, пищевые сети животных, нейронные сети, компьютерные и он-лайн сети | Сетевые структуры различной природы: социальные он-лайн и офф-лайн сети | Выявление имплицитных он-лайн сообществ, связанных веб-документов и сайтов | Выявление сообществ в социальных (в частности, он-лайн) социальных сетях | Анализ структуры комплексных, неполных и неточных данных: медицинских, биологических, он-лайн |
| Реальные области применения | Выявление степени совпадения символического состава информации | | Выявление плагиата, связанных публикаций | Выявление сообществ в сетях различной природы | | Выявление связанных сайтов | Распространение информации | - |
| Время выполнения | 30000000 (150 Гб) 10,5 дней | 500 Мб – 1 Гб 1 час | 2000 документов 3 минуты | 56276 вершин 42 минуты | 15000 вершин 6295801 ребер 15,18 с | | | |
| Метод | Анализ символического состава информации, кластеризация | | Анализ связей, контент-анализ | Кластеризация | | | | |
| Программное обеспечение | | sif | TREC | | Metis | SDOC | | |
| Необходимые ресурсы | Изначальное множество сравниваемых объектов | | Пользовательский запрос, поисковая система | Граф (визуализированные объекты и связи между ними) | | Список кластеризуемых объектов и описание их характеристик | | |
| Степень вовлеченности (затратности, участия) исследователя | Выбор размера «шингла» | Выбор размера «якоря», порога совпадения | Определение параметров, формулировка запроса | Построение изначального графа | | Описание объектов | Распространение «виртуальных» чернил | Описание объектов |
| | | | | | | | | |
| Недостатки | Отсутствие описания промежуточных стадий. | Невозможность содержательного анализа. | Трудоемкость, отсутствие критериев выбора параметров | Значительные временные затраты на реализацию | Отсутствие описания способов реализации стадий алгоритма | Путаница в терминологии | Теоретический | Трудоемкость характер |

**Заключение**

Бесспорно, сегодня Интернет-среда стремительно растет, и практика диктует запрос на всё новые, более эффективные методы обработки больших массивов информации. Для поиска информации и принятия решения эксперту, аналитику или простому пользователю уже недостаточно использовать поисковые машины. Для наиболее эффективного решения поставленных задач и поиска необходимой информации нужны новые методы, которые позволили бы пользователю (в широком смысле этого слова) при минимальных усилиях получить необходимую информацию.

Для настоящей работы были подобраны современные методы анализа и выделения сетевых сообществ, представляющие два различных направления (математическое и социолингвистическое). На первом этапе методы были классифицированы по направлениям принадлежности Затем были подробно разобраны типичные представители каждого направления. На финальной стадии методы были распределены, исходя их задач, ими решаемых. Такое подход отвечает практическим запросам – ведь на практике исследователь исходит именно из практической задачи и подбирает методы и процедуры для её решения, а не наоборот.

В заключение необходимо отметить, что данная работа является начальным этапом для более глубоких и масштабных теоретико-методологических исследований и их практического применения. В дальнейшем планируется разработать методологию (с подробным описанием алгоритма в целом и конкретных процедур) выделения сообщества (скорее всего, экспертного), учитывающую достоинства, недостатки и целевую направленность проанализированных методов. Продолжением теоретических изысканий может стать онтологическая основа методов, расширение и углубление произведенной классификации.