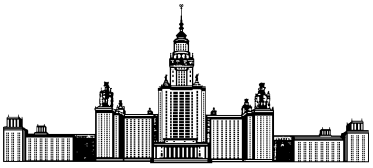
Московский государственный университет имени М. В. Ломоносова



Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Дипломная работа

Восстановление связей в научном рубрикаторе на основе кластеризации гетерогенной сети

Выполнил:

студент 5 курса 517 группы

Борисов Михаил Викторович

Научный руководитель:

к.ф-м.н., доцент

Майсуразде Арчил Ивериевич

Москва, 2014

Оглавление

1 Введение

1.1. Наукометрия как предметная область

Наукометрия — научная дисциплина, изучающая развитие науки путем измерения различных связанных с научной деятельностью показателей [38, 12]. Считается, что впервые термин «наукометрия» был введен В.В. Налимовым в 1969 году в монографии [42] «Наукометрия: Изучение науки как информационного процесса». Результаты наукометрии активно используются при оценке результативности работы научных институтов, в том числе, при решении вопросов об объемах и эффективности их финансирования.

В своей деятельности наукометрия во многом опирается на достижения библиометрии — более прикладной дисциплины, занимающейся количественной и качественной оценкой влияния научных публикаций. Основными методами библиометрии являются цитатный анализ и контент-анализ. Цитатный анализ использует достижения прикладной математики для построения социологического описания изучаемого научного сообщества: рассматривается информация, которую можно извлечь из связей, создаваемых за счет цитирования научных работ. Контент-анализ применяет лингвистические инструменты для решения задач по извлечению смысла публикаций. Типичным примером задачи, решаемой в рамках контент-анализа, является получение краткого описания публикации в виде набора характерных ключевых фраз.

Наукометрическими задачами занимаются эксперты аналитических центров. Одной из массовых задач, стоящих перед ними, является задача рубрикации, то есть определения принадлежности научных статей заранее заданным областям знаний. В зависимости от выбранной модели, структуры данных и методики классификации, а так же набора библиометрических инструментов, задача рубрикации статей может быть поставлена по-разному.

1.2. Подходы к решению задач рубрикации научных публикаций

Рассмотрим, какие подходы к классификации научных статей по предопределенным областям знаний описаны в литературе.

Исторически самый первый способ классификации — ручное формирование группой экспертов правил отнесения к тем или иным классам. Этот подход дорогостоящий и дает невысокую обобщающую способность, многие признают его устаревшим [30].

Наше внимание будет в первую очередь уделено методам, основанным на машинном обучении. Детальный обзор таких методов приведен в [30]. Можно заметить, что методы делятся на подгруппы по разным характеристикам:

- По структуре множества классов: плоская или иерархическая. Как правило, плоская структура сочетается с довольно грубым делением на области знаний (которые могут быть как непересекающимися, так и пересекающимися), а иерархическая — с подробной онтологией.

- По виду отображения: назначение документа одному или нескольким классам (в том числе отдельно рассматривается задача с одним классом: документ ставится в соответствие строго либо этому классу, либо его дополнению).

- Документо-центричные или категорие-центричные: найти все категории для данного документа или все документы для данной категории.

- «Жесткое» соответствие или «мягкое» (ранжированное) — ставится ли документу в соответствие определенная категория, или только задается порядок (оценки) на множестве категорий. «Жесткое» соответствие полезно для полностью автоматических систем принятия решений, в то время как при полуавтоматической работе эксперты могут быть удовлетворены ответом в «мягкой» форме.

2 Общая задача восстановления связей в научном рубрикаторе

В данной главе описывается модель данных, используемая в работе аналитическим центром. Формулируются содержательные и формальные постановки решаемых в работе задач. Рассматриваются представленные в литературе подходы к решению близких задач.

2.1. Выбор модели данных

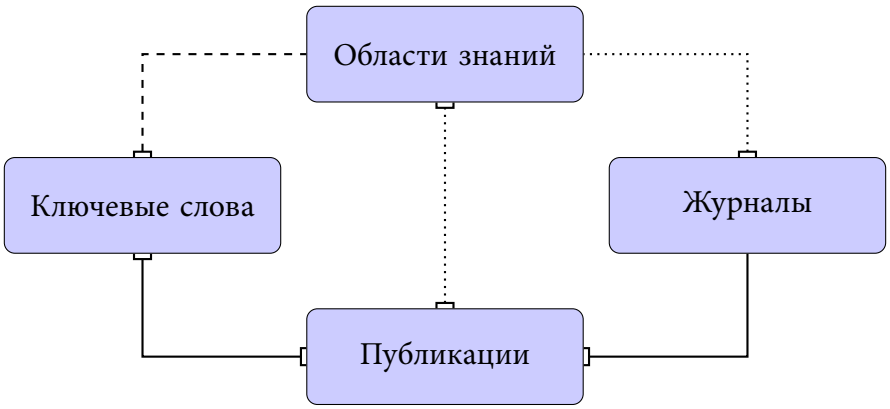


Рис. 2.1: Структура модели данных.

В работе рассматривается ситуация, когда эксперты аналитического центра предпочитают использовать модель на основе реляционных СУБД. Это решение обусловлено тем, что РСУБД давно зарекомендовали себя как эффективный подход к хранению и обработке больших объемов связанных между собой данных. В рамках выбранной модели, данные рассматриваются как гетерогенный граф, в вершинах которого находятся единицы анализа, а ребра соответствуют связям между ними. Гетерогенность графа означает, что его вершины соответствуют множествам объектов различных типов. Такая структу-

ра естественным образом отображается в схему базы данных посредством механизмов внешних ключей и развязочных таблиц.

Единицами анализа в выбранном представлении данных являются публикации, речевые маркеры (ключевые слова), издания (напр. научные журналы) и области знаний (рис. 2.1). Публикации связаны с изданиями, в которых они опубликованы (одна публикация — не более, чем в одном издании). Для каждой публикации известен набор ключевых слов, ее описывающих. Для некоторых ключевых слов вручную проставлены их однозначные соответствия областям знаний. Дополнительно можно рассматривать соответствия публикаций и изданий областям знаний. Такие соответствия являются по сути композициями промежуточных соответвий: изданий — публикациям, публикаций — ключевым словам, и ключевых слов — областям знаний.

Следует отметить, что выбор РСУБД как средства представления данных оставляет открытым вопрос об ограничениях на характер связей («один к многим», «многие ко многим») между объектами. И если в некоторых случаях связи задаются естественным образом (напр. одна статья может быть опубликована только в одном журнале), то при определении соответствий объектов областям знаний есть выбор между двумя типами моделей. В «мягкой» модели допустимо отнесение одного объекта к взвешенному набору областей знаний, в то время как в «жесткой» модели один объект не может быть отнесен более чем к одной области знаний. Преимуществом «мягкой» модели является то, что она лучше отображает характер неоднозначности соответствия, но в то же время, подобные неоднозначности зашумляют набор данных, понижая информативность. «Жесткость» модели при этом можно рассматривать как фактор ее регуляризации. Из соображений регуляризации и легкости интерпретации представления данных, была выбрана «жесткая» модель.

Таким образом, используемая в работе модель является:

- Плоской — на множествах объектов не задано иерархии,
- Трехуровневой — публикации связаны только с ключевыми словами, ключевые слова далее связаны только с областями знаний.
- Жесткой — для каждого ключевого слова его связь с областью знаний находится в одном из трех состояний: либо задана конкретная область знаний, либо о связи еще ничего не известно, либо принято решение отказа от классификации.

2.2. Постановка содержательной задачи

Рассматривается описанная выше плоская трехуровневая жесткая модель научного рубрикатора. Связи между статьями и ключевыми словами, а также между статьями и журналами рассматриваются как полные и заранее заданные (внесены экспертами). Соответствие между ключевыми словами и областями знаний задано не полностью: ввиду большого размера множества возможных связей, размечена только часть пар.

Задача рубрикации состоит в том, чтобы для каждой публикации указать ее принадлежность к нескольким рубрикам, либо сообщить об отказе от классификации. При наличии связей между ключевыми словами и областями знаний, либо между изданиями и областями знаний, задачу можно было бы решить путем транзитивного замыкания. Таким образом, можно свести исходную задачу к двум подзадачам:

- Использование существующих связей «ключевое слово — область знаний». При этом как создаются новые связи, так и используются уже заданные экспертами.
- Использование связей соответствия журналов областям науки (требуется вывести через прочие связи). При этом по тем же соображениям, как и для связей между ключевыми словами и областями знаний, рассматриваются «жесткие» соответствия, т.е. журнал может быть отнесен к не более чем одной области. Существенное отличие этой задачи от предыдущей состоит в том, что связи создаются «с нуля».
- Композиция двух предыдущих подходов: исследовать, как использование дополнительной информации позволяет повысить качество.

Нелишне отметить, что среди возможных ответов следует рассматривать и отказ от классификации, так как не все из имеющихся ключевых слов можно отнести к какой-либо конкретной области знаний.

2.3. Формальная постановка задачи

Введем обозначения, которые будут использованы ниже в данной работе. Обозначим единичные объекты (см. табл. 2.1) как P, K, F, J. Множества объектов типа X обозначим как V_X, а множества пар объектов $\langle X, Y \rangle$ — как E_{XY}.

Таблица 2.1: Обозначения

Тип	Один объект	Множество объектов
Публикация	P	V_P
Ключевое слово	K	V_K
Область знаний	F	V_F
Издание	J	V_J
Связь «публикация — ключевое слово»	$\langle P, K \rangle$	$E_{PK} \subset V_P \times V_K$
Связь «ключевое слово — область знаний»	$\langle K, F \rangle$	$E_{KF} \subset V_K \times V_F$

Запишем формальные постановки подзадач.

Подзадача 1. Пусть даны множества объектов V_P, V_K, V_F, множества пар E_{PK}, E^0_{KF}, а так же функционал качества $F(E_{KF}|V_P, V_K, V_F, E_{PK})$. Требуется найти решение — множество связей E^*_{KF}, минимизирующее некоторый заданный функционал F, т.е.

$$E^*_{KF} = \arg \min_{E_{KF} \subset V_K \times V_F} F(E_{KF}|V_P, V_K, V_F, E_{PK}, E^0_{KF}).$$

Подзадача 2. Пусть даны множества объектов V_P, V_K, V_F, V_J, множества пар E_{PK}, E_{KF}, E_{PJ}, а так же функционал качества $F(E_{JF}|V_P, V_K, V_F, V_J, E_{PK}, E_{KF}, E_{PJ})$. Требуется найти решение — множество связей E^*_{JF}, минимизирующее некоторый заданный функционал F, т.е.

$$E^*_{JF} = \arg \min_{E_{JF} \subset V_J \times V_F} F(E_{JF}|V_P, V_K, V_F, V_J, E_{PK}, E_{KF}, E_{PJ}).$$

Отметим, что вид функционала качества намеренно оставлен свободным, чтобы путем подбора такого функционала выбирать метод решения подзадачи. Кроме того, в практических целях задача скорее всего будет ослаблена до нахождения приблизительного решения ввиду того, что непосредственное решение подобной задачи оптимизации крайне трудоемко.

2.4. Обзор литературы

В данном разделе рассмотрены подходы и алгоритмы, решающие задачи, сходные с поставленной в рамках данной работы.

2.4.1. Восстановление связей

В научной литературе весьма широко освещена проблема предсказания и восстановления в реляционных данных. К сожалению, в этих работах авторы как правило рассматривают гомогенные сети, что не позволяет непосредственно применить описанные ими методы для решения нашей задачи. При обзоре литературы использовались материалы обзора [8].

Распространенным подходом является переход к задаче классификации пар узлов («потенциальных ребер») на два класса: «присутствие ребра» и «отсутствие ребра» [26, 25, 17, 18]. Однако, стоит отметить, что данный подход затруднительно применить при наличии структурных ограничений на связи, например, связей вида «один к одному» или «один ко многим».

В [26] рассматриваются проблемы оценки качества алгоритмов восстановления связей. Авторы указывают на тот факт, что распределение прецедентов по классам является сильно несимметричным, так как с ростом размера графа число существующих связей для объекта растет как правило не более чем линейно, в то время как количество отсутствующих связей растет квадратично. Авторы описывают постановку задачи обнаружения аномальных, т.е. статистически неправдоподобных связей.

В [13] изучаются эффекты смещенности признаковых описаний и статистической зависимости между признаками. Авторы исследуют извлечение реляционных признаков, то есть таких признаков объекта, которые зависят от признаков соседних (в смысле связей между объектами) объектов. Авторы описывают два явления: *концентрированную связность* (*concentrated linkage*), то есть ситуацию, когда многие объекты связаны с одним общим, и *реляционную автокорреляцию* (*relational autocorrelation*), что означает существенную равномерность значений признаков среди соседей одного объекта.

В [25] авторы рассматривают задачу предсказания связей-цитирований между документами на основе базы данных CiteSeer. Ссылаясь на [13], акцентируют внимание на том, что обязательно требуется учитывать реляционную структуру данных. Авторы применяют логистическую регрессию как алгоритм распознавания потенциальных связей и изучают влияние используемых признаков и структуры набора данных.

В [9, 37] рассматривается задача восстановления связей с точки зрения традиционных задач машинного обучения.

В [15] авторы предлагают использовать реляционную марковскую сеть — разновидность марковских сетей для реляционных моделей данных.

В [17] авторы описывают исследование прогнозирования возникновения связей в социальных сетях с течением времени. В качестве данных используется граф соавторств публикаций на ресурсе arXiv. Авторы задаются вопросом, насколько эффективно можно моделировать поведение сети, учитывая только внутренние факторы (привязанные к сети), но не имея информации о внешних. Предсказание связей между объектами (авторами) рассматривается как задача ранжирования пар объектов. Для этого авторы рассматривают различные функции, имеющие смысл оценки близости объектов в парах.

В [18] рассматриваются подходы обучения с учителем (supervised) и без учителя (unsupervised) при предсказании новых связей, авторами делается вывод о превосходстве методов с учителем.

В [19] рассматривается задача восстановления связей, аналогичная исследуемой в рамках данной работы. Авторы переходят от задачи предсказания связей в графе к задаче предсказания связей разных типов. Предлагаемый метод также сравнивается с несколькими альтернативными методами. Авторы рассматривают тройки, состоящие из пары объектов разных типов и типа связи между ними. Используя данные о сходстве таких троек и ссылаясь на методы [16, 31] частичного обучения [41], авторы предлают свой алгоритм предсказания связей.

2.4.2. Ассоциативные правила

Если рассматривать общую задачу, рассматриваемую в данной работе, как задачу кластеризации долей гетерогенной сети, то можно заметить определенное сходство с областью, изучаемой при помощи механизма ассоциативных правил. В самом деле, так как мера сходства объектов определяется ссылками, близки оказываются понятия «входят в один кластер» и «часто встречаются вместе», в то время как последнее является типичной областью применения ассоциативных правил.

На основании [10, 1], а так же более ранних публикаций [2, 7, 5, 40] можно получить представление о классической постановке задачи поиска ассоциативных правил. Рассматривается двудольный граф, в котором одна доля — объекты, другая — транзакции. Каждой транзакции сопоставлен некоторый набор объектов. Рассматривается задача выделения групп объектов, которые часто вместе участвуют в транзакциях.

Более формально, вводятся следующие понятия. Пусть I — множество рассматриваемых объектов. Поддержка $\sigma(X)$ набора объектов $X \subseteq I$ — количество транзакций, в которые весь набор входит как подмножество. Набор X считается популярным, если

$\sigma(X) > min\_sup$. Ассоциативное правило — выражение $A \Rightarrow B$, где $A \subset I$ и $B \subset I$ — наборы объектов, причем $A \cap B \neq \varnothing$. Для ассоциативного правила вводятся: поддержка — $\sigma(A \Rightarrow B) = \sigma(A \cup B)$ и достоверность — $\text{conf}(A \Rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$. Поддержка правила характеризует, насколько часто встречаются наборы объектов, на основе которых можно вывести данное ассоциативное правило, в то время как достоверность характеризует «вероятность» того, что данное следствие имеет место. Аналогично наборам данных, рассматриваются ассоциативные правила, чья поддержка превышает порог *min_conf*. Далее задача поиска ассоциативных правил разбивается на две подзадачи:

1. Поиск популярных наборов объектов, т.е. удовлетворяющих нижнему порогу поддержки *min_sup*.

2. Порождение достоверных ассоциативных правил (т.е. с достоверностью, превышающей нижний порог *min_conf*, из наборов данных, полученных на шаге 1).

При этом, наибольшие трудности вызывает первый шаг, так как требуется существенное сокращение размерности пространства перебора относительно полного (множества всех подмножеств множества объектов), имеющего экспоненциальную относительно количества объектов размерность.

Первым предложенным методом поиска популярных наборов объектов является APriori [3]. Этот алгоритм основан на важном свойстве антимонотонности ассоциативных правил, которое заключается в том, что поддержка множества не может превосходить наименьшей поддержки среди подмножеств этого множества. Алгоритм использует это свойство для поиска всех популярные наборов алгоритмом поиска в ширину, использующим в качестве базы тривиальные одноэлементные множества объектов. Ввиду не очень высокой эффективности (т.к. алгоритм должен в явном виде сгенерировать наборы-кандидаты в популярные наборы), данный алгоритм как правило представляет преимущественно академический интерес.

Альтернативным, более новым, методом является FP-growth [21], основанный на структуре данных, названной FP-дерево (frequent-pattern tree). Для построения этого дерева, объекты упорядочиваются по убыванию их поддержки; в соответствие каждому объекту ставится уровень в дереве. Таким образом, любой набор объектов может быть представлен в виде некоторого пути от корня дерева вниз. Эффективность хранения достигается за счет того, что для часто встречающихся объектов вместо множества их экземпляров хранятся количества. Далее приводится алгоритм построения популярных наборов на основе построенного FP-дерева.

Помимо традиционного механизма ассоциативных правил, так же рассматриваются обобщенные ассоциативные правила [35, 32]. При этом множество объектов представляется в виде иерархической таксономии, что позволяет находить взаимоотношения как между отдельными объектами, так и между их группами.

Одним из расширений механизма ассоциативных правил является механизм последовательных шаблонов (sequential patterns) [4, 36]. В этом случае, кроме отношения «встречаются вместе» рассматривается так же и порядок, в котором объекты были включены в набор.

К сожалению, несмотря на то, что ассоциатиные правила предоставляют достаточно мощный теоретический и практический аппарат, их непосредственное применение при решении данной задачи затруднительно.

2.4.3. Кластеризация графов

Одним из подходов к распространению меток в графе может быть кластеризация графа, с последующим отнесением кластеров ключевых слов к одной области знаний. Такой подход операется на гипотезу о том, что отмеченные экспертами вручную области знаний хорошо коррелируют с особенностями внутренней структуры графа.

Исследуем существующие подходы кластеризации графов, существенную помощь в этом помощь оказали обзоры в книгах [20, 39, 23].

В первую очередь можно выделить алгоритмы кластеризации гомогенных графов. В таких моделях рассматриваются графы отношений между объектами одного типа [20, 11].

Существуют методы [20, 37], основанные на вероятностной формулировке. По структуре графа строится байесовская сеть, на которой применяются методы работы с графическими моделями, например, belief propagation.

Среди существующих подходов кластеризации, применительно к данной задаче интерес представляет бикластеризация [20, 6, 33]. Бикластеризацией называется одновременная кластеризация узлов каждой из компонент двудольного графа с сохранением структуры связей между полученными кластерами. Распространенным подходом для бикластеризации двудольного графа является спектральный метод кластеризации [22, 33, 34]. Его идея заключается в том, что матрица смежности графа с выраженными кластерами может быть преобразована к блочной форме (где каждый блок соответствует определенному кластеру, и значения внутри блока различаются меньше, чем значе-

ния в разных блоках) перестановкой строк и столбцов. Далее ищется факторизация $\|Z_{N\times M} - R_{N\times K}B_{K\times L}C_{L\times M}\| \to \min$ (где N и M - размеры долей в исходном графе, K и L - размеры кластерных описаний этих долей соответственно).

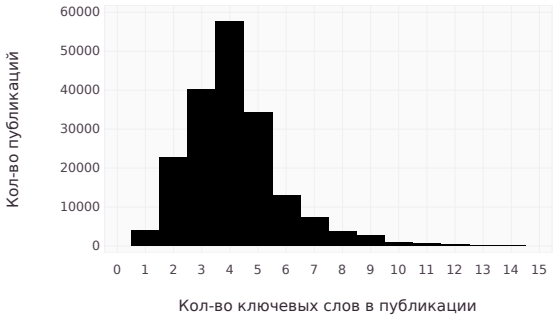
2.5. Анализ набора исходных данных

Имея доступ к примеру исходного набора данных, изучим несложные количественные характеристики модели, которые позволят более объективно оценивать применимость и эффективность различных подходов и алгоритмов.

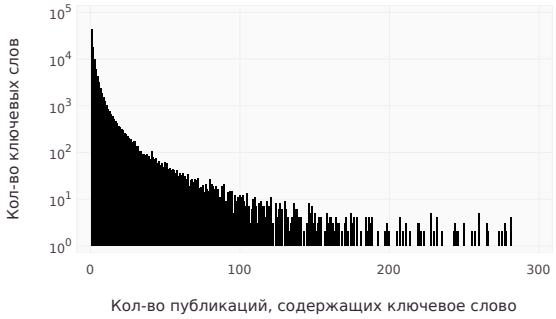
Таблица 2.2: Количественное описание набора данных

Единица анализа	Количество
Области знаний	48
Публикации	364221
Ключевые слова	335645
Связи «публикация — ключевое слово»	1147427
Экспертные связи «ключевое слово — область знаний»	474

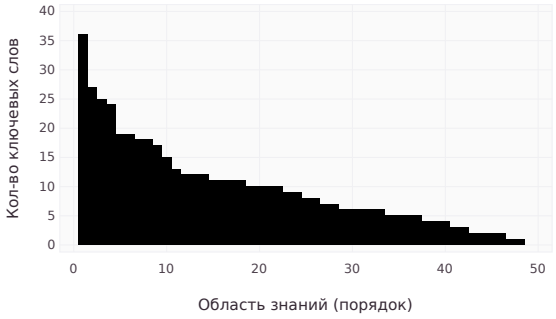
В табл. 2.2 представлены количества записей, составляющих входные данные. График 2.2a показывает распределение количеств ключевых слов, сопоставленных различным публикациям. График 2.2d показывает распределение количеств областей знаний, транзитивно соответствующих публикациям (считаем, что область знаний F транзитивно соответствует публикации P, если существует такое ключевое слово K, что $\exists \langle P, K\rangle \in E_{PK}$ И $\exists \langle K, F\rangle \in E_{KF}$).



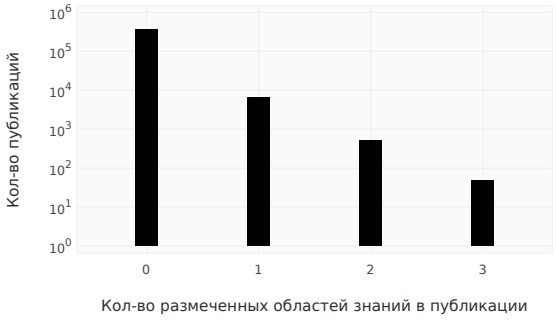
(a) Количества ключевых слов на публикацию



(b) Количества публикаций на ключевое слово



(c) Количество экспертных ключевых слов на область знаний



(d) Количество областей знаний на публикацию (при транзитивном соответствии)

Рис. 2.2: Статистические характеристики набора данных

3 Обогащение неполных связей посредством классификации объектов

В данной подзадаче рассматриваются только связи между публикациями, ключевыми словами и областями знаний (журналы исключаются из модели). Диаграмму сокращенной модели можно наблюдать на рис. 3.1.

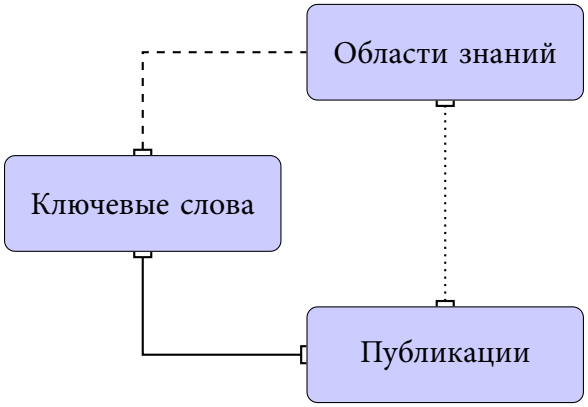


Рис. 3.1: Структура модели данных подзадачи 1.

3.1. Формальная постановка задачи

Целью данной части работы является обнаружение новых соответствий ключевых слов областям знаний. В терминах задач классификации, можно разбить эту задачу на две подзадачи:

- Формирование признакового описания для каждого ключевого слова.

- Решение задачи классификации по полученному признаковому описанию и исходным соответствиям ключевых слов областям знаний в качестве набора прецедентов.

3.2. Формирование признакового описания

Существенной частью задания является построение «плоского» признакового описания по графовому описанию исходной задачи. При построении признакового описания применимы разные подходы, обладающие различными преимуществами и недостатками.

3.2.1. Признаки на основе совместных частот

Достаточно простым алгоритмом является алгоритм, основанный на частотных описаниях ключевых слов по их совместной встречаемости с областями знаний. То есть, для каждой публикации рассматриваются неструктурированные наборы ключевых слов и областей знаний (область знания включается, если хоть какое-либо из заранее размеченных ключевых слов, входящих в набор ключевых слов, соответствующих данной публикации, соответствует данной области знания). Воспользуемся терминологией, близкой к области извлечения ассоциативных правил.

Рассмотрим множества речевых маркеров I_K и областей науки I_F. Составим из них общее множество объектов $I = I_K \cup I_F$. По множеству документов, составим транзакции (каждый документ порождает не более одной транзакции). В транзакцию включим все речевые маркеры, входящие в данный документ, и все области знаний, к которым относится хотя бы один речевой маркер. Если ни один из речевых маркеров ни отнесен ни к одной из областей знаний, документ транзакции не порождает. К полученным транзакциям можно применять методы извлечения ассоциативного правила, для извлечения транзакций вида $i_k \Rightarrow i_f(i_k \in I_K, i_f \in I_F)$. Формальное описание алгоритма представлено в алг. 1.

Таким образом, после первого этапа получены признаковые описания в виде описаний вида (область знаний → количество совместных упоминаний). При этом, так как в одном документе могут быть несколько ключевых слов, сопоставленных одной области знаний, допустимы варианты:

- каждая область знаний учитывается не более одного раза

- каждая область знаний учитывается столько раз, сколько соответствующих ей ключевых слов содержатся в документе.

Алгоритм 1 Построение признакового описания на основе частот

Вход: $\text{frequency}(D, f)$ — функция подсчета частоты области знаний f в документе D
Выход: $K \in \mathbb{R}^{+}_{N_k \times N_f}$ — Матрица признаков

```
K_{k,f} ← 0        ∀k, f
Для каждого документа D :
    Для каждого k ∈ keywords(D) :
        Для каждого f ∈ fields_of_science(D) :
            K_{k,f} ← K_{k,f} + frequency(D, f)
```

Недостатком данного метода является то, что он способен учитывать только локальную структуру данных. Это обусловлено тем, что учитываются только непосредственные связи — при данном подходе невозможно передать информацию транзитивно через несколько объектов (например, между двумя документами, имеющими общее ключевое слово).

3.2.2. Признаки на основе расстояний

Если предыдущий метод рассматривает только непосредственные связи и различает их по частоте, альтернативный подход заключается в том, чтобы вычислять длины минимальных путей в графе, связывающих ключевые слова с областями знаний. Связями первого порядка считаются непосредственные связи ключевых слов с областями знаний. Связями второго порядка — связи между ключевым словом и областями знаний, соответствующими ключевым словам, содержащимся в документах, содержащих текущее ключевое слово, и т.д.

При расчетах удобно хранить положительные целые расстояния, либо 0 как флаг «бесконечного расстояния». После расчета, все положительные расстояния заменяются на обратные: таким образом, значения однородно кодируют меру сходства (т.к. 0 является обратным к «бесконечности»).

Пусть $I \subset I_k \times I_f$ — набор обучающих ассоциаций ключевых слов с областями знаний. Введем две матрицы $K_{k,f}$ — расстояний от ключевого слова k до области знаний f, и $D_{d,f}$

— расстояний от документа d до области знаний f. В качестве базы примем $D = (0)$ и

$$K_{k,f} = \begin{cases} 1 & \text{если } (k,f) \in I, \\ 0 & \text{иначе} \end{cases}.$$

На каждой итерации будем осуществлять переход от цепочек длиной i к цепочкам длиной $i+1$. Для этого, сначала обновляем матрицу D по матрице K, записывая новые расстояния между документами и областями знаний, после чего аналогичным образом обновляем матрицу K по матрице D, записывая новые расстояния между ключевыми словами и областями знаний. По построению, данный алгоритм является адаптацией стандартного алгоритма обхода в ширину для выбранной структуры данных. Полезным свойством является то, что расстояния могут меняться лишь однажды — с бесконечности до конечного значения, и не обновляются позже (что обеспечивается невзвешенностью ребер графа). Формальное описание алгоритма представлено в алг. 2.

3.3. Выбор решающего правила

После составления признакового описания ключевых слов, требуется сформулировать функцию, принимающую решение об отнесении ключевого слова к той или иной области знаний. Допустимыми ответами для этой функции считаются утвердительный ответ об отнесении ключевого слова к одной определенной области знаний, либо отказ от классификации. Отказ от классификации может быть вызван как нехваткой знаний о ключевом слове (в случае, когда оно участвует в критически малом числе связей), так и неоднозначность ключевого слова. Неоднозначными считаются ключевые слова, имеющие слишком общее значение — такое, что оно не ползволяет отнести их к каким-либо конкретным областям знаний. Примерами таких слов могут быть EXPERIMENTAL, EXPLICIT, REAL, NUMBER.

Полезным свойством выбранных признаковых описаний является то, что число признаков равно числу классов и каждый признак уже указывает на близость к соответствующему классу, причем, чем сильнее соответствие, тем больше значение данного признака. Такое свойство подсказывает очевидного кандидата на решающее правило: сопоставить ключевому слову класс, которому соответствует наибольшее значение признака сходства. В случае, если наибольная мера сходства равна нулю (что значит, что абсолютно все меры сходства равны нулю) следует принять решение об отказе от классифика-

Алгоритм 2 Построение признакового описания на основе расстояний

Вход: N — максимальная глубина цепочки
Вход: $I = \{(k,f)\}$ — набор обучающих ассоциаций ключевых слов с областями знаний
Выход: $K \in \mathbb{R}^{+}_{N_k \times N_f}$ — Матрица признаков

$K_{k,f} \leftarrow 0 \qquad \forall k,f$
$D_{d,f} \leftarrow 0 \qquad \forall d,f$
Для каждого $(k,f) \in I$ **:**
    $K_{k,f} \leftarrow 1$
Цикл $i \leftarrow 1 \ldots N$ **:**
    Для каждого k — ключевое слово **:**
        Для каждого f — область знаний **:**
            Если $K_{k,f} = i$ **тогда**
                Для каждого d — документ **:**
                    Если $D_{d,f} = 0$ **тогда**
                        $D_{d,f} \leftarrow i + 1$
    Для каждого d — документ **:**
        Для каждого f — область знаний **:**
            Если $D_{d,f} = i$ **тогда**
                Для каждого k — ключевое слово **:**
                    Если $K_{k,f} = 0$ **тогда**
                        $K_{d,f} \leftarrow i + 1$
Для каждого k — ключевое слово **:**
    Для каждого f — область знаний **:**
        Если $K_{k,f} \neq 0$ **тогда**
            $K_{k,f} \leftarrow 1/K_{k,f}$

ции. Также, следует учитывать проблему неоднозначности классификации посредством выбора максимального значения, если несколько признаков имеют равные наибольшие положительные значения. В данном случае требуется вводить иерархию классов — например, всегда выбирать класс с наименьшим номером: такое правило позволяет стабилизировать результат работы алгоритма, но не несет никакой содержательной информации, так как классы могут быть заранее упорядочены произвольным образом. Альтернативой является обогащать признаковое описание в надежде на то, что другие признаки позволят ликвидировать неоднозначность.

3.4. Методология оценки качества решения

Существующая структура исходных данных затрудняет их разбиение на обучение и контроль. Поэтому предлагается использовать следующие функционалы:

- Leave-One-Out: доля неправильных восстановлений связи «ключевое слово — область знаний» при ее исключении из обучающего набора данных. Следует обратить внимание, что исключается как можно меньший объем данных, а именно информация о единственной связи, которую предполагается восстановить. Все прочие связи сохраняются как есть.
- Полнота — доля вновь классифицированных ключевых слов. Предполагается, что лучшее решение обладает более высокой полнотой.
- Однозначность — распределение количества областей знаний, к которым может быть транзитивно отнесена публикация. Могут рассматриваться как скалярные статистики этого распределения, так и его гистограмма целиком. Предполагается, что каждая публикация как правило относится к одной области знаний, редко — к двум, и крайне редко — к трем и более (см. рис. 2.2d), поэтому поощряется соответствующая форма распределения, полученного после классификации.

3.5. Исследуемые алгоритмы

Ниже, под SVM понимается многоклассовый SVM классификатор, построенный из N одноклассовых по стратегии «один-против-всех».

Таблица 3.1: Алгоритмы классификации

Алгоритм	Признаковое описание	Решающее правило
Freq-Unique-Argmax	частотное описание без повторений	arg max
Freq-All-Argmax	частотное описание с повторениями	arg max
Freq-Unique-SVM	частотное описание без повторений	SVM
Freq-All-SVM	частотное описание с повторениями	SVM
Dist-Restrict-Argmax	расстояния с учетом только обучающей выборки ключевых слов	arg max
Dist-Full-Argmax	расстояния с учетом всех ключевых слов	arg max
Dist-Restrict-SVM	расстояния с учетом только обучающей выборки ключевых слов	SVM
Dist-Full-SVM	расстояния с учетом всех ключевых слов	SVM
Mixed-Restrict-All-SVM	частоты с повторениями и расстояния с учетом только обучающей выборки ключевых слов	SVM
Mixed-Restrict-Unique-SVM	частоты без повторений и расстояния с учетом только обучающей выборки ключевых слов	SVM

3.6. Проведенные эксперименты

При программировании экспериментов использовалась система компьютерной алгебры Julia [14].

На основе данных из транзакций, полученных описанным выше методом, сформированы признаковые описания: каждому ключевому слову сопоставлены частотные описания его совместной встречаемости с областями знаний. Итого из исходных 335645 ключевых слов получено 18240 непустых признаковых описания, из них 474 размеченных (т.е. пригодных для использования в качестве обучающей выборки).

Была проведена оценка алгоритмов по методике Leave-One-Out: для каждой пары «ключевое слово — область знаний»:

1. составлены частотные признаковые описания каждого ключевого слова, не использующие информацию из выбранной для исключения пары;

2. исключенному ключевому слову сопоставлена предсказанная область знаний.

В таблице 3.2 приведены результаты замера показателей алгоритма на тесте Leave-One-Out. Пусть алгоритм дал N_{total} ответов всего, из них: $N_{correct}$ верных, N_{wrong} неверных,

Таблица 3.2: LOO тестирование алгоритмов

Алгоритм	% Accuracy	% Error	% Ambiguous	% Reject
Freq-Unique-Argmax	19,2	58,4	22,4	53,8
Freq-All-Argmax	20,0	58,0	21,9	53,8
Freq-Unique-SVM	19,2	65,8	15,0	53,8
Freq-All-SVM	18,7	63,5	17,8	53,8
Dist-Restrict-Argmax	14,7	54,3	31,0	47,1
Dist-Full-Argmax	**23,2**	**44,6**	**32,2**	**6,9**
Dist-Restrict-SVM	8,1	82,2	9,7	47,1
Dist-Full-SVM	7,0	68,3	24,7	6,9
Mixed-Restrict-All-SVM	6,5	33,4	60,0	38,7
Mixed-Restict-Unique-SVM	6,3	33,4	60,2	38,7

$N_{ambiguous}$ классификаций как неоднозначные и N_{reject} отказов от классификации. В таком случае,

$$\begin{aligned}
\text{Accuracy} &= \frac{N_{correct}}{N_{total} - N_{reject}} \\
\text{Error} &= \frac{N_{wrong}}{N_{total} - N_{reject}} \\
\text{Ambiguous} &= \frac{N_{ambiguous}}{N_{total} - N_{reject}} \\
\text{Reject} &= \frac{N_{reject}}{N_{total}}
\end{aligned}$$

Результаты тестирования показали, что как правило, решающее правило на основе SVM дает худшие результаты, чем argmax. В некоторых случаях SVM показал существенно худшие результаты. Кроме того, можно заметить, что частотные признаковые описания лучше подходят для применения SVM, чем основанные на расстояниях.

Лучшие результаты показали: признаковое описание, основанное на расстояниях, с решающим правилом argmax, либо частотное описание с решающим правилом argmax, независимо от способа подсчета. Стоит заметить, что частотное описание является существенно менее вычислительно затратным.

3.7. Выводы

Эксперименты не позволили найти алгоритм, существенно превосходящий по качеству «наивный». К сложалению, все алгоритмы демонстрируют удручающе низкое качество классификации.

4 Восстановление связей при помощи кластеризации графа

Подход, альтернативный классификации ключевых слов, заключается в кластеризации вершин графа с последующим сопоставлением кластеров конкретным областям знаний. При «оптимистичном» сценарии, в результате кластеризации ключевые слова будут разбиты на относительно небольшие группы с близким смыслом, после чего эксперт может перейти от разметки отдельных слов, к пакетной разметке групп слов, что позволит повысить производительность ручного труда эксперта. Однако, возможен и «пессимистичный» сценарий, когда результат кластеризации слабо коррелирует с целевым отображением, т.е. когда большинство кластеров состоят преимущественно из слов, относящихся к различным областям знаний.

4.1. Постановка задачи

Требуется найти такое разбиение на кластеры исходного множества, в котором, во-вервых, число кластеров было бы минимально, во-вторых, преобладали бы связи внутри кластеров, а не между кластерами, в-третьих, каждый кластер должен соответствовать наименьшему числу областей знаний. Данная постановка соответствует постановке задачи бикластеризации графа, что определяет выбор методов ее решения.

4.2. Алгоритм

Был выбран алгоритм спектральной бикластеризации [6]. Этот алгоритм имеет один существенный параметр: количество целевых кластеров. Так как в нашей постановке задачи, количество кластеров заранее неизвестно (требуется группировать любые похожие объекты, но не разбивающиеся априори на некоторое заранее заданное число клас-

сов). При выборе значения этого параметра, требуется найти оптимальное значение. Из предположения, что присутствуют ключевые слова, относящиеся к каждой области знаний, разбиения на число кластеров меньше, чем число областей знаний, очевидно, не имеют смысла. Чем меньше целевых классов выбрано, тем меньше времени требуется для поиска разбиения и тем менее однозначным является разбиение на кластеры с точки зрения отображения ключевых слов в области знаний. С другой стороны, разумное ограничение сверху на число кластеров задает число размеченных ключевых слов: при слишком мелком разбиении, многие кластеры не будут содержать размеченных ключевых слов, что значит, что их затруднительно применять для автоматического анализа.

Оценим величины значений, исходя из размеров набора данных. Так как в наборе знаний присутствуют 48 областей знаний и 474 размеченных ключевых слова, выбор менее 50 либо более 500 кластеров является заведомо непрактичным. При этом, средний размер кластера приблизительно равен от 700 (для разбиения на 500 кластеров), до 7000 (для разбиения на 50 кластеров) ключевых слов, что в любом случае является весьма большим для человеческого анализа количеством.

Алгоритм 3 Бикластеризация связей мужду публикациями и ключевыми словами

Вход: $M_{N_p \times N_k}$ — матрица смежности публикаций и ключевых слов
Вход: N_c — количество целевых кластеров
Выход: $C^p_{N_p \times N_c}$ — матрица разметки публикаций
Выход: $C^k_{N_k \times N_c}$ — матрица разметки ключевых слов
1: $C^p, C^k \leftarrow \text{spectral\_coclustering}(M, N_c)$

4.3. Проведенные эксперименты

При программировании экспериментов использовались среда программирования Python [27], а так же библиотеки численных вычислений NumPy [24], SciPy [29], scikit-learn [28]

Были построены распределения размеров кластеров, полученных в результате бикластеризации связей между публикациями и ключевыми словами для количества кластеров от 50 до 625 с шагом 50. В таблице 4.1 приведено распределение размеров полученных кластеров в областях знаний. Столбцам таблицы соответствует количество областей знаний, в ячейках — доля кластеров, попавших в этот диапазон, в процентах. Со-

ответственно таблице, также построен график изменения долей диапазонов этого значения при увеличении числа кластеров.

Эксперименты показали, что не более 20% кластеров могут быть однозначно отнесены к какому-либо из ключевых слов. При этом, начиная с приблизительно 400 кластеров не менее 65% кластеров соответствуют не более чем 5 областям знаний. Существенного повышения однозначности разбиения при увеличении числа кластеров сверх 400 не наблюдается, что подтверждает гипотезу о верхнем пределе размера разбиения множества на кластеры. Кроме того, наблюдаемые при большом числе кластеров затраты оперативной памяти и процессорного времени делают использование разбиений на такое число кластеров еще менее практичным.

Таблица 4.1: Распределение однозначности отображений кластеров

Число кластеров	0	1	2–3	4–5	6–7	8–10	11–15	15+
50	0	0	0	0	0	0	14	86
75	0	0	0	0	0	5	46	48
100	0	0	0	0	2	33	45	20
125	0	0	1	2	14	31	39	11
150	0	0	0	7	17	38	28	8
175	0	0	2	13	23	36	20	3
200	0	0	5	17	31	27	16	2
225	0	1	8	24	27	24	12	1
250	0	1	11	28	28	18	10	0
275	0	1	14	28	27	20	6	0
300	0	4	19	27	26	16	5	0
325	0	4	20	30	27	12	4	0
350	1	3	24	34	22	9	2	0
375	2	7	24	34	17	12	1	0
400	2	7	29	34	15	9	1	0
425	2	9	33	31	15	6	1	0
450	4	11	32	30	13	6	1	0
500	3	10	37	30	13	4	0	0
525	5	11	41	26	11	2	0	0
575	7	13	41	24	9	3	0	0
600	6	18	40	25	7	2	0	0
625	6	16	41	24	8	1	0	0

4.4. Выводы

Подход бикластеризации графа показал себя непрактичным для целей автоматической разметки, однако, данный подход можно применить для помощи в ручной работе аналитикам, так как он позволяет эффективно сократить неоднозначность в определении конкретного класса. К его недостаткам можно также отнести неспособность выявлять многозначные слова и высокую вычислительну стоимость.

5 Заключение

В рамках выполнения этой работы была сформулирована формальная задача восстановления связей в научном рубрикаторе. Был проведен обзор литературы в областях восстановления связей в сетях, ассоциативных правил, бикластеризации сетей, а также частичного обучения. Обзор литературы показал, что не существует готовых методов решения поставленной задачи.

Были сформулированы способ построения алгоритмов классификации, решающих поставленную задачу, а так же методика их оценки. На основе полученного подхода, составлен набор исследуемых алгоритмов. Исследуемые алгоритмы реализованы и протестированы на имеющемся наборе данных. На основе полученных измерений, выбран алгоритм, обеспечивающий наилучшее качество распознавания.

Была сформулирована постановка задачи с т.з. кластеризации графа. Были изучены методы кластеризации графа. Проведены эксперименты с реализацией бикластеризации графа. Изучена структура результата в зависимости от числа выделяемых кластеров.

Литература

1. *Adamo J.* Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms. — Springer New York, 2001.
2. *Agrawal R.*, *Imielinski T.*, *Swami A.* Mining Association Rules between Sets of Items in Large Databases // IN: PROCEEDINGS OF THE 1993 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, WASHINGTON DC (USA). — 1993. — С. 207—216.
3. *Agrawal R.*, *Srikant R.* Fast Algorithms for Mining Association Rules. — 1994.
4. *Agrawal R.*, *Srikant R.* Mining Sequential Patterns //. — 1995. — С. 3—14.
5. *Chen M.-S.*, *Han J.*, *Yu P. S.* Data mining: an overview from a database perspective // Knowledge and data Engineering, IEEE Transactions on. — 1996. — Т. 8, № 6. — С. 866—883.
6. *Dhillon I. S.* Co-clustering documents and words using Bipartite Spectral Graph Partitioning. — 2001.
7. Fast Discovery of Association Rules. / R. Agrawal [и др.] // Advances in knowledge discovery and data mining. — 1996. — Т. 12. — С. 307—328.
8. *Getoor L.*, *Diehl C. P.* Link Mining: A Survey // SigKDD Explorations Special Issue on Link Mining. — 2005.
9. *Getoor L.*, *Taskar B.* Introduction to statistical relational learning. — MIT press, 2007.
10. *Han J.*, *Kamber M.*, *Pei J.* Data Mining: Concepts and Techniques: Concepts and Techniques. — Elsevier Science, 2011. — (The Morgan Kaufmann Series in Data Management Systems).
11. *Hirano S.*, *Tsumoto S.* Hierarchical Clustering of Non-Euclidean Relational Data Using Indiscernibility-Level. // RSKT. Т. 5009 / под ред. G. Wang [и др.]. — Springer, 15 мая 2008. — С. 332—339. — (Lecture Notes in Computer Science).

12. *Hood W. W.*, *Wilson C. S.* The literature of bibliometrics, scientometrics, and informetrics // Scientometrics. — 2001. — Т. 52, № 2. — С. 291—314.

13. *Jensen D.*, *Neville J.* Linkage and autocorrelation cause feature selection bias in relational learning // ICML. Т. 2. — Citeseer. 2002. — С. 259—266.

14. Julia: A Fast Dynamic Language for Technical Computing / J. Bezanson [и др.] // CoRR. — 2012. — Т. abs/1209.5145.

15. Label and link prediction in relational data / B. Taskar [и др.] // Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data. — Citeseer. 2003.

16. Learning with Local and Global Consistency. / D. Zhou [и др.] // NIPS. Т. 16. — 2003. — С. 321—328.

17. *Liben-Nowell D.*, *Kleinberg J.* The link-prediction problem for social networks // Journal of the American society for information science and technology. — 2007. — Т. 58, № 7. — С. 1019—1031.

18. *Lichtenwalter R. N.*, *Lussier J. T.*, *Chawla N. V.* New perspectives and methods in link prediction // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2010. — С. 243—252.

19. Link propagation: A fast semi-supervised learning algorithm for link prediction. / H. Kashima [и др.] // SDM. Т. 9. — SIAM. 2009. — С. 1099—1110.

20. *Long B.*, *Zhang Z.*, *Yu P.* Relational Data Clustering: Models, Algorithms, and Applications. — Chapman & Hall/CRC, 2009. — (Chapman & Hall/CRC Data mining and knowledge discovery series).

21. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach / J. Han [и др.]. — 2004.

22. *Ng A. Y.*, *Jordan M. I.*, *Weiss Y.* On Spectral Clustering1 Analysis and an algorithm // Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press. — 2001. — Т. 14. — С. 849—856.

23. *Noel S.*, *Raghavan V.*, *Chu C.-H. H.* Document clustering, visualization, and retrieval via link mining // Clustering and Information Retrieval. — Springer, 2004. — С. 161—193.

24. *Oliphant T. E.* A Guide to NumPy. Т. 1. — Trelgol Publishing USA, 2006.

25. *Popescul A.*, *Ungar L. H.* Statistical relational learning for link prediction // IJCAI workshop on learning statistical models from relational data. Т. 2003. — Citeseer. 2003.

26. *Rattigan M. J.*, *Jensen D.* The case for anomalous link discovery // SIGKDD Explor. — 2005.

27. *Rossum G. van* [и др.] Python programming language. — 1991–.

28. Scikit-learn: Machine learning in Python / F. Pedregosa [и др.] // The Journal of Machine Learning Research. — 2011. — Т. 12. — С. 2825—2830.

29. SciPy: Open source scientific tools for Python / E. Jones, T. Oliphant, P. Peterson [и др.]. — 2001–.

30. *Sebastiani F.* Machine learning in automated text categorization // ACM computing surveys (CSUR). — 2002. — Т. 34, № 1. — С. 1—47.

31. Semi-supervised learning using gaussian fields and harmonic functions / X. Zhu, Z. Ghahramani, J. Lafferty [и др.] // ICML. Т. 3. — 2003. — С. 912—919.

32. *Silverstein C.*, *BRIN S.*, *MOTWANI R.* Beyond Market Baskets: Generalizing Association Rules To Dependence Rules. — 1998.

33. Spectral Biclustering of Microarray Cancer Data: Co-clustering Genes and Conditions / Y. Kluger [и др.] // Genome Research. — 2003. — Т. 13. — С. 703—716.

34. Spectral clustering for multi-type relational data / B. Long [и др.] // In ICML. — 2006. — С. 585—592.

35. *Srikant R.*, *Agrawal R.* Mining Generalized Association Rules //. — 1995. — С. 407—419.

36. *Srikant R.*, *Agrawal R.* Mining Sequential Patterns: Generalizations and Performance Improvements // Research Report RJ 9994, IBM Almaden Research. — 1995.

37. *Taskar B.*, *Segal E.*, *Koller D.* Probabilistic classification and clustering in relational data // International Joint Conference on Artificial Intelligence. Т. 17. — LAWRENCE ERLBAUM ASSOCIATES LTD. 2001. — С. 870—878.

38. *Van Raan A. F.* Scientometrics: State-of-the-art // Scientometrics. — 1997. — Т. 38, № 1. — С. 205—218.

39. *Yu P.*, *Han J.*, *Faloutsos C.* Link Mining: Models, Algorithms, and Applications. — Springer, 2010.

40. *Zaki M. J.*, *Ogihara M.* Theoretical foundations of association rules // 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. — Citeseer. 1998. — С. 71—78.

41. *Zhu X.* Semi-supervised learning literature survey // Computer Science, University of Wisconsin-Madison. — 2006. — Т. 2. — С. 3.

42. *Налимов В. В.*, *Мульченко З. М.* Наукометрия: Изучение развития науки как информационного процесса. — Изд-во «Наука», 1969.