На правах рукописи

Сивоголовко Елена Владимировна

ОЦЕНКА КАЧЕСТВА КЛАСТЕРИЗАЦИИ В ЗАДАЧАХ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Специальность 05.13.11 —
Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Санкт-Петербург
2014

Работа выполнена на кафедре системного программирования математико-механического факультета федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет»

Научный руководитель: доктор физико-математических наук, профессор Новиков Борис Асенович

Официальные оппоненты: Баранов Сергей Николаевич, доктор физико-математических наук, профессор, заведующий лабораторией Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук

Ступников Сергей Александрович, кандидат технических наук, старший научный сотрудник Федерального государственного бюджетного учреждения науки Институт Проблем информатики Российской академии наук

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Южно-Уральский Государственный Университет» (национальный исследовательский университет)

Защита диссертации состоится “15” мая 2014 года в 15 часов на заседании диссертационного совета Д 002.087.01 при Федеральном государственном бюджетном учреждении науки Институте системного программирования Российской академии наук по адресу: 109004, Москва, ул. Александра Солженицына, д. 25.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института системного программирования Российской академии наук.

Автореферат разослан “12” апреля 2014 года.

Ученый секретарь
диссертационного совета
кандидат физ.-мат. наук, /Зеленов С. В./

Общая характеристика работы

Актуальность темы. Интеллектуальный анализ данных (англ. Data Mining) — собирательное название, используемое для обозначения совокупности методов обнаружения в исследуемых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Кластеризация(или же кластерный анализ) является одной из центральных областей интеллектуального анализа данных. Её задачей является выделение в исходном множестве некоторой, заранее неизвестной, структуры кластеров. Термин кластер понимается в интеллектуальном анализе данных как объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определёнными свойствами.

Кластеризация широко используется как в качестве отдельного инструмент анализа, так и как один из этапов предварительной обработки данных перед использованием других аналитических методов (например, перед классификацией или поиском ассоциативных правил). Кажущаяся простота постановки задачи, возможность выявления скрытых группировок элементов данных как на множестве в целом, так и внутри каждого кластера по отдельности, а так же несомненное улучшение восприятия аналитиками данных, разбитых на кластеры, — всё это делает методы кластерного анализа широко используемыми в самых различных областях. Кластеризации подвергаются данные по экономике и социологии, тексты, новостные потоки, блоги, генетические последовательности, изображения, данные социальных сетей, медицинские и биологические показатели.

При получении какого-либо результата у исследователя, аналитика или разработчика естественным образом возникает вопрос о его качестве. Точно так же вопрос о качестве возникает и после проведения кластерного анализа данных. Оценка качества несомненно важна для всего процесса кластеризации, потому что без неё полученная структура кластеров не может быть сочтена достаточно достоверной для того, чтобы делать из неё определённые выводы или проводить дальнейший анализ на её основе. Широко распространённым способом оценки качества кластеризации является проверка, производимая аналитиком вручную, например, с помощью визуализации полученной кластерной структуры. Однако, это является удобным далеко не всегда, а в случае больших объёмов данных или же данных с высокой размерностью, такая проверка достаточно затруднительна. Альтернативой методам визуальной оценки качества являются автоматические методы оценки качества кластеризации. Они могут быть рассмотрены как функции от полученной кластерной структуры и исходного множества. В литературе методы данной группы носят названия индексы или метрики. Именно такой спектр методов является объектом исследования в данной работе.

Разнообразные задачи кластерного анализа данных традиционно привлекают к себе внимание исследователей на протяжении нескольких десятилетий и до текущего момента не потеряли своей актуальности. Поскольку оценка качества результатов неотделима от общего процесса кластеризации, разработка, сравнение или повышение эффективности автоматических методов оценки качества кластеризации являются актуальными областями исследований.

Цель работы. Целью работы является исследование методов оценки качества кластеризации, выявление их сильных и слабых сторон и создание специальной библиотеки индексов оценки качества кластеризации, которой смогут пользоваться аналитики разных направлений.

Для достижения поставленной цели были выделены следующие задачи:

- Сравнение уже имеющихся индексов оценки качества кластеризации и определение их эффективности в зависимости от используемого алгоритма кластеризации.
- Выявление зависимости качества кластеризации от качества исходных данных.
- Разработка модели качества кластеризации, учитывающей формализованное представление пользователя о желаемой кластерной структуре.

Основные результаты. В диссертации получены следующие результаты:

1. Предложен метод выбора индексов оценки качества в зависимости от используемых алгоритмов кластеризации.

2. Предложен метод выбора алгоритмов кластеризации в зависимости от ожидаемого уровня качества данных.

3. Построена семантическая модель оценки качества кластеризации, основанная на сравнении формализованного представления пользователя о результатах кластеризации с полученной структурой кластеров.

4. Представлен индекс оценки качества кластеризации, построенный на основе разработанной модели.

5. Разработана библиотека индексов оценки качества кластеризации, содержащая реализации методов, рассмотренных в ходе проведённых исследований.

Научная новизна. В работе предложена модель оценки качества кластеризации, отражающая семантику полученного результата кластеризации. Под семантикой в данном случае понимается выраженные в терминах предложенной модели представления пользователя о результатах кластеризации.

В соответствии с этой моделью введены понятия размерностей качества кластеризации, и представлен метод их измерения в терминах модели Resource Description Framework(RDF).

Выяснено влияние измерений качества исходных данных на качество результатов кластеризации, и сделаны выводы об использовании алгоритмов кластеризации в случае соответствующих ошибок данных.

Сформулированы требования к методам оценки качества кластеризации в зависимости от используемого алгоритма.

Теоретическая ценность и практическая значимость. Теоретическую ценность работы составляет предложенная семантическая модель оценки качества кластеризации, позволяющая исследователям и аналитикам оценивать валидность кластерной структуры с семантической точки зрения.

В рамках данной работы разработан прототип библиотеки методов оценки качества кластеризации, включающий в себя реализацию одиннадцати относительных индексов и четырёх внешних индексов оценки качества. Библиотека написана на языке Java и является совместимой с платформой Weka.

Апробация работы. Основные результаты диссертации докладывались на семинаре Московской Секции ACM SIGMOD; на семинарах группы исследования методов организации информации при лаборатории исследования операций НИММ СПбГУ, а также на следующих конференциях:

1. Симпозиуме Молодых Учёных SYRCoDIS, 2009;

2. Докторском консорциуме объединенных конфференций EDBT/ICDT "EDBT/ICDT PhD Workshop" 2012;

3. Международной конференции "Advanced Databases and Information Systems" 2012;

4. Международной Балтийской Конференции по Базам Данных и Информационным Системам "Baltic Conference on Databases and Information Systems" 2012;

Публикации. Основное содержание диссертации опубликовано в восьми научных статьях, список которых приведён в конце автореферата. Две из этих статей находятся в изданиях, рекомендуемых ВАК для публикации основных результатов диссертаций на соискание учёной степени кандидата наук.

Основными работами по теме диссертации являются

1. **Сивоголовко Е.** Методы оценки качества четкой кластеризации. // Компьютерные инструменты в образовании — Тверь, 2011 — Вып. 4 (96) — С. 14-31.

2. **Sivogolovko E.** The Influence of Data Quality on Clustering Outcomes. // Frontiers in Artificial Intelligence and Applications — IOS Press, 2012 — Vol. 249 — P. 95–105.

3. **Sivogolovko, E., Thalheim, B.** Semantic approach to cluster validity notion. // Advances in Databases and Information Systems / Ed. by Tadeusz Morzy, Theo Harder, Robert Wrembel. — Springer Berlin Heidelberg, 2012 — Vol. 186 — P. 229–239.

Структура и объем диссертации. Диссертация состоит из пяти глав, включая введение; заключения и списка литературы, содержащего восемьдесят семь названий. Общий объём диссертации составляет девяносто две страницы.

Содержание работы

Глава 1, обзорная посвящена обзору предшествовавших результатов, относящихся к теме диссертации, в ней обсуждаются мотивировки решаемых задач и даются необходимые определения и обозначения.

В разделах 1.1-1.2 приводится понятие качества кластеризации и индексов оценки этого качества. «Оптимальная» структура кластеров зачастую определяется как результат работы алгоритма кластеризации, наилучшим образом соответствующий скрытому во множестве разбиению (Halkidi M.). Однако, это определение не является единственным, также в литературе можно встретить и следующую трактовку: адекватность структуры кластеров соответствует степени, в которой кластерная структура представляет реальную информацию о данных или же способность полученной структуры отражать внутренние взаимодействия данных (Dubes R. C.).

Для формализации понятия качества используются индексы (в некоторых источниках их также называются методами или метриками) оценки качества кластеризации. Обычно, они представляют собой функции от кластерной структуры : $F(\mathcal{C}) \to \mathbb{R}_+$, в некоторых случаях и от структуры, и от исходного множества: $F(\mathcal{C}, \mathcal{D}) \to \mathbb{R}_+$.

В данной работе рассматриваются исключительно индексы для оценки качества чёткой неиерархической кластеризации. Термин «чёткая» используется для описания кластерной структуры, кластеры внутри которой не пересекаются. В свою очередь, термин «неиерархическая» обозначает структуры, не обладающие вложенностью кластеров, порождающей иерархию. В разделе 1.3 приведены описания алгоритмов чёткой неиерархической кластеризации, которые использовались в работе. Таких алгоритмов четыре: K-Means, X-Means, Father First и DBScan. По принципу своей работы первые три из них являются центроориентированными, а последний — плотностным. Именно алгоритмы этих классов в настоящее время используются наиболее широко.

Основное различие для методов данных семейств заключается в определении того, что такое кластер. В центроориентированных алгоритмах элемент x принадлежит кластеру C, если расстояние между данным элементом и центром кластера C является минимальным из всех расстояний между x и центрами построенных кластеров. В свою очередь, понятие кластера для DBScan вводится следующим образом: пусть ϵ -окрестность элемента x — это $N_\epsilon(x) = \{y \in X | d(x,y) \leq \epsilon\}$, элемент x кластеризуемого множества является элементом кластерного «ядра», если его ϵ -окрестность содержит не менее чем $minPts$ элементов. Элемент y является достижимым в смысле плотности (*density-reachable*) из элемента ядра x если существует конечная последовательность элементов ядра между x и y, такая что каждый следующий ее элемент принадлежит к ϵ-окрестности предыдущего. Элемент y является связанным в смысле плотности (*is density-connective*) с элементом x, если они оба достижимы в смысле плотности из одного и того же элемента ядра. Соответственно, кластером для алгоритма DBScan является компонента связности из достижимых по плотности элементов. В разделе кратко описаны схемы работы каждого из алгоритмов из используемых алгоритмов.

Раздел 1.4 посвящён описанию текущих подходов к измерению качества кластеризации. В нем подробно описаны три основных подхода к оценке качества кластерной структуры: внутренний, внешний и относительный. Приведены примеры индексов оценки для каждого из них. Основной упор сделан на индексы относительного подхода, так как именно они являются объектом исследования данной диссертации. Сделаны выводы о проблемах, существующих в этой области.

Глава 2 посвящена исследованию относительных индексов оценки качества для чёткой неиерархической кластеризации. Приводится сравнение выбранных индексов при работе с плотностными и разбивающими алгоритмами.

В разделе 2.1 обосновывается необходимость проведения исследования подобного рода. Ставятся задачи, во-первых, определить насколько те или иные индексы эффективны для оценки результатов кластеризации, полученных с помощью алгоритмов различной природы, и, во-вторых, разработать методы тестирования индексов качества. Под эффективностью в данных разделах понимается способность индекса оценки выбирать в качестве оптимальной структуры кластеров ту, которая наиболее полно отражает имеющиеся во множестве отношения близости между элементами.

Раздел 2.2 посвящён описанию предложенного автором процесса тестирования индексов оценки качества и используемых при этом инструментов и множеств. Были использованы два алгоритма кластеризации разной природы: центроориентированный K-Means и плотностной DBSCan. В качестве тестовых множеств данных были взяты четыре множества с разной степенью кластеризуемости: данные без структуры кластеров, данные с вложенными кластерами, данные с плохоотделимыми кластерами, и данные с хорошо от-

делимыми кластерами.

В разделах 2.3-2.4 описаны эксперименты по проверке эффективности индексов оценки качества. Эффективность алгоритма определялась как количество совпадений результата работы алгоритма с определённой вручную «правильной» структурой кластеров. Ниже приведены сделанные из этих экспериментов выводы:

- ни один индекс не может определить, что во множестве отсутствует структура кластеров, однако, из того, что показания группы индексов значительно расходятся друг с другом, можно сделать выводы о том, что используемое множество данных плохо поддаётся кластеризации.
- индексы, которые при построения модели качества структуры учитывают только компактность кластеров, а отделимость не учитывают, менее эффективны.
- если для кластеризации используются разбивающие алгоритмы, такие как K-Means, то нет особой разницы, какой из описанных индексов брать для оценки качества, не включая индексы, описанные в предыдущем пункте.
- Если для кластеризации используется DBScan, то лучше использовать индексы, учитывающие геометрическую структуру кластеров (CDbw) и измеряющие компактность и отделимость в терминах средних расстояний между элементами кластеров (Silhouette).
- Для получения объективного результата лучше пользоваться не одним конкретным индексом, а их совокупностью; нельзя утверждать, что один конкретный индекс всегда будет показывать оптимальный результат.
- Лучшие индексы в порядке уменьшения точности оценки: 1) CDbw 2)индексы силуэта 3)Dunn и DB.

Индексы, используемые в этой части работы, легли в основу разработанной библиотеки индексов оценки качества кластеризации.

Глава 3 посвящена изучению влияния качества данных на качество кластеризации.

В разделе 3.1 производится формулировка задачи, решаемой в этой главе. Чаще всего подразумевается, что кластеризуемое множество не содержит ошибок. В том случае, если наличие ошибок оговаривается отдельно, они все идут под одним общим названием — «шум». При этом, само понятие «шума» никак не связывается ни с одним из понятий, используемых в области качества данных, в которой существует достаточно подробная классификация вероятных ошибок. Возникает определённая разорванность между этими двумя областями. Таким образом, если у аналитика имеются предположения о качестве имеющихся данных, как он сможет использовать их при выборе алгоритма кластеризации?

В этой части работы исследовалось влияние различных измерений (или сторон) качества данных на качество кластеризации. Поставленная здесь основная цель была не столько подтвердить, что снижение качества данных влечёт за собой снижение качества кластеризации (это представляется в достаточной степени очевидным), сколько понять какие из измерений качества имеют наибольшее негативное влияние, ошибки какого рода наиболее «опасны» для кластеризации. Для этого были взяты четыре измерения качества данных, пять множеств данных и четыре алгоритма кластеризации. Сначала проводилась кластеризация выбранных тестовых множеств как идеальных по качеству, и измерялось качество кластеризации в этом - идеальном - случае. После этого для каждого множества данных на каждом измерении качества моделировались три уровня качества: высокий, средний и низкий в соответствии с представленным процессом моделирования ошибок данных. Затем проводилась кластеризация искусственно ухудшенных данных, оценивалось качество этой кластеризации и его ухудшение в каждом случае по сравнению с качеством кластеризации для идеального множества данных.

В разделе 3.2 даны определения четырёх основных измерений качества данных, которые использовались в работе: точности, полноты, согласованности и своевременности. Точность рассматривается как количество элементов с верными значениями атрибутов. Полнота рассматривается как мера того, что все атрибуты данных попали в базу. Согласованность определяют как однородность элементов — отсутствие конфликтов внутри и между множествами данных. Своевременность определяют через два понятия: изменчивость (volatility) и срок действия(currency). Изменчивость характеризует период времени между одним изменением в реальном мире и следующим за ним изменением, которое делает реальные данные не валидными. Срок действия определяет «возраст» данных, используемых для получения информации. В рамках данной работы, срок действия данных был принят неопределённо большим, и своевременность определялась исключительно как изменчивость. Приводятся процентные соотношения неверных данных, характерных для низкого, среднего и высокого уровня качества по каждому измерению.

Раздел 3.3 содержит описание предложенного автором процесса моделирования ошибок в данных, в соответствии с определениями, данными в предыдущем разделе.

Разделы 3.4-3.5 содержат описание используемого тестового окружения: тестовых множеств данных и алгоритмов, применяемого метода подбора параметров для кластеризации, проводимых экспериментов и методов статистической оценки, используемых для определения статистической значимости полученных результатов.

Раздел 3.6 посвящён итогам исследования, описанного в этой главе. Было продемонстрировано, что первые три из рассматриваемых сторон качества: точность, полнота и согласованность при ухудшении имеют статистически значимое и отрицательное влияние на качество кластеризации, в то время как

четвёртое – своевременность – имеет статистически значимый отрицательный эффект при использовании алгоритмов, аргументы которых получены в результате некоторой оценки «идеального» множества данных и нуждаются в пересчёте при добавлении во множество новых элементов или изменении уже существующих (как это происходит с параметром $minPts$ в DBScan). Была построена взаимосвязь между качеством данных и качеством кластеризации и даны рекомендации по использованию различных алгоритмов кластеризации, в зависимости от того, какого рода ошибки в данных ожидает встретить аналитик. Список сторон качества данных в порядке уменьшения влияния на качество кластеризации выглядит следующим образом:

1. Согласованность.
2. Полнота.
3. Точность.
4. Своевременность.

Наиболее устойчивые к ошибкам алгоритмы кластеризации для соответствующих сторон качества данных:

1. Согласованность: нет алгоритмов.
2. Полнота: FarthestFirst и XMeans.
3. Точность: KMeans и XMeans.
4. Своевременность: XMeans.

В **главе 4** представлен, разработанный в диссертации, семантический подход к определению качества кластеризации. Задача этой главы формулируется следующим образом: имеется некое представление пользователя о структуре кластеров и кластерная структура, построенная по его запросу, нужно определить насколько данная структура соответствует этому представлению. Соответственно, для её достижения требуется: формализовать представление пользователя, формализовать информацию, полученную при кластеризации и сравнить эти объекты. Полученный результат будет являться семантической оценкой качества кластеризации. Описанию этих действий посвящены следующие разделы этой главы.

В разделе 4.1 освещается суть поставленной задачи: в настоящее время индексы качества кластеризации не учитывают представления пользователя о желаемом результате кластеризации и семантику кластерной структуры в процессе оценки качества. Исходя из этого, необходим подход, который принимал бы эти понятия во внимание.

В разделе 4.2 даётся определение синтаксического качества кластеризации как качества, которое вычисляется относительно внутренних свойств полученной структуры кластеров или же свойств исходного множества. К таким свойствам, например, относятся, среднее расстояние между элементами в исследуемом множестве, среднее расстояние между элементами кластеров и межкластерное расстояние. Структура кластеров и множество данных в этом случае рассматриваются исключительно как множества векторов в пространстве соответствующей размерности.

В разделе 4.3 вводится определение семантического качества кластеризации, как степени соответствия между теми сведениями, которые можно извлечь из полученной структуры кластеров и формализованным представлением пользователя об итогах кластеризации. Тем самым, фокус оценки качества смещается с определений некоторых свойств множества векторов на семантику его содержания. Рассмотрим простейшую задачу кластеризации на примере множества данных IRIS. С точки зрения пользователя она может быть поставлена следующим образом: «Имеется множество данных из трёх кластеров, нужно найти по одному ключевому представителю для каждого кластера, ожидается, что дисперсия по каждому кластеру не будет превышать 0.2. » На этом небольшом примере легко видеть, что все представление пользователя об итогах кластеризации можно разделить на три части: предварительная информация о множестве, необходимые цели и ожидания. Описанное представления пользователя формализовать в виде набора предикатов: «Объект» – « Предикат» – «Субъект». Для построения семантической модели качества кластеризации в данной работе будет рассматриваться только та информация, которую можно формализовать в виде обозначенной выше тройки.

Далее в этом же разделе вводится модель семантического качества, состоящая из четырёх основных частей:

1. $\mathcal{D}$ — формализованная информация, которая содержится во множестве данных и может быть извлечена с помощью кластеризации.

2. $\mathcal{C}$ — формализованная информация, которая может быть извлечена из построенной кластерной модели.

3. $\mathcal{U}$ — формализованная информация, получаемая со стороны пользователя. В нашем представлении она состоит из трёх разных частей: $\mathcal{U} = \mathcal{U}_K \cup \mathcal{U}_G \cup \mathcal{U}_E$, здесь $\mathcal{U}_K$ представляет собой текущее знание пользователя о множестве данных, $\mathcal{U}_G$ — концептуальное представление пользователя о том, что ему желательно получить в результате кластеризации, и $\mathcal{U}_E$ — пользовательские «ожидания», спектр информации, которая не является необходимой, но которую пользователь может ожидать и объяснить, исходя из имеющихся у него представлений и сведений о процессе кластеризации для данного множества.

4. $\mathcal{V}$ — условия действия – общие правила, которые должны соблюдаться во всех элементах модели.

Далее понятие семантического качества формализуется в терминах измерений.

Раздел 4.4 посвящён измерению семантического качества с помощью концепции Resource Description Framework(RDF). В нем задача оценки семантического качества сводится к задаче сравнения двух RDF-графов, и вводится специальный индекс для их сравнения на основе контрастной модели Тверски. В разделе предложен способ отображения в RDF модели семантического качества из раздела 4.3, а также мера сходства RDF-графов, которая может быть использована для оценки семантического качества кластеризации.

Пусть RDF граф G — это множество триплетов (s, p, o), таких что $s \in \mathbf{U} \cup \mathbf{B}, p \in \mathbf{U}$ и $o \in \mathbf{U} \cup \mathbf{L} \cup \mathbf{B}$. Здесь $\mathbf{U}$ — множество идентификаторов ресурсов, $\mathbf{L}$ — множество литералов и $\mathbf{B}$ — множество пустых вершин.

Определим функцию описания $DF : \mathbf{G} \times 2^{\mathbf{U}} \to \mathbf{G}$, где $\mathbf{G}$ есть множество всех RDF графов, как $DF(G, W) = \{(s, p, o) | (s, p, o) \in G \& s \in W\}$. Неформально, $DF(G, W)$ возвращает описание всех ресурсов, принадлежащих ко множеству W.

Определим следующие множества сравнения для графа G и множества ресурсов X, входящих в G:

- $P(X) = \{p | (s, p, o) \in DF(G, X) \& s \in X\}$ – множество свойств (предикатов).
- $D(X) = \{o | (s, p, o) \in DF(G, X) \& s \in X \& o \in \mathbf{L}\}$ – домен.
- $O(X) = \{o | (s, p, o) \in DF(G, X) \& s \in X \& o \in \mathbf{U}\}$ – множество объектов.
- $T(X) = \{(p, o) | (s, p, o) \in DF(G, X) \& s \in X\}$ – множество типов.

Пусть X_1 — множество всех ресурсов, используемых в графе G_1, и X_2 — множество всех ресурсов в графе G_2 соответственно. Для ресурса $x \in X_1 \cap X_2$ обозначим за $P_1(x), D_1(x), O_1(x), T_1(x)$ множества сравнения между ним и графом G_1. Аналогично для того же ресурса x и графа G_2. Тогда определим меру сходства между графами G_1 и G_2 следующим образом:

$$GraphSim(G_1, G_2) = \sum_{x \in X_1 \cup X_2} (SetSim(P_1(x), P_2(x)) + SetSim(D_1(x), D_2(x)) + SetSim(O_1(x), O_2(x)) + SetSim(T_1(x), T_2(x)))$$

Здесь $SetSim$ является мерой схожести двух множеств и задается с помощью модели Тверски:

$$SetSim(A, B) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}, \alpha, \beta \geq 0$$

Нормализованный вариант меры схожести графов:

$$NGraphSim(G_1, G_2) = \frac{1}{4}\frac{1}{|X_1 \cup X_2|} * GraphSim(G_1, G_2)$$

Таким образом, по значению $NGraphSim(C, U)$ можно судить о степени сходства полученной кластерной структуры с формализованным пользовательским представлением о ней, что, в соответствии с введённым определением, является семантическим качеством кластеризации.

В **заключении** кратко сформулированы основные результаты диссертации, к которым относятся:

1. Сравнение эффективности методов оценки качества кластеризации для разных классов алгоритмов кластеризации и рекомендации по выбору индекса оценки качества относительно используемого алгоритма.

2. Определение влияния различных сторон качества данных на качество кластеризации. Ранжирование сторон качества по степени влияния на кластеризацию. Рекомендации по выбору алгоритмов кластеризации в зависимости от ожидаемого уровня качества данных.

3. Понятие семантического качества кластеризации и его формализация в терминах отдельных измерений.

4. Индекс оценки семантического качества кластеризации, основанный на сравнении RDF графов.

5. Создание библиотеки индексов оценки качества, в которую вошли четыре внешних индекса для чёткой кластеризации и одиннадцать относительных индексов для чёткой кластеризации. Библиотека написана на языке Java и может быть интегрирована в систему интеллектуального анализа данных Weka.

Работы автора по теме диссертации

Публикации автора в изданиях из списка ВАК и изданиях, находящихся в международных индексах цитирования

1. **Сивоголовко Е.** Методы оценки качества четкой кластеризации. // Компьютерные инструменты в образовании — Тверь, 2011 — Вып. 4 (96) — С. 14-31.

2. **Сивоголовко Е.** Методы обобщенной кластеризации при анализе социальных сетей. // Программные продукты и системы — 2011 — Вып. 4 — С. 98–101.

3. **Sivogolovko E., Novikov B.** Validating cluster structures in data mining tasks. // Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012 — ACM, 2012 — P. 245–250.

4. **Sivogolovko, E., Thalheim, B.** Semantic approach to cluster validity notion. // Advances in Databases and Information Systems / Ed. by Tadeusz Morzy, Theo Harder, Robert Wrembel. — Springer Berlin Heidelberg, 2012 — Vol. 186 — P. 229–239.

Публикации автора в других изданиях

5. **Sivogolovko E.** Cluster validity in high-dimensional spaces. // ADBIS (local proceedings) — 2008 — P. 172–176.

6. **Sivogolovko E.** Clustering and ranking of image search results. // Proceedings of the 6th Spring Young Researchers' Colloquium on Databases and Information Systems (SYRCoDIS-2009), Saint-Petersburg, Russia, May 28-29, 2009. –— Saint-Petersburg, 2009.

7. **Sivogolovko E.** Evaluation of impact of data quality on clustering with syntactic cluster validity methods. — Kiel, 2011 — P. 37 — (Institut f ür Informatik der Christian-Albrechts-Universität zu Kiel, Bericht Nr. 1107, ISSN 2192-6247).

8. **Simoes J., Kiseleva J., Sivogolovko E., Novikov B.** Exploring influence and interests among users within social networks. // Computational Social Networks — Springer London, 2012 — P. 177–206.

9. **Sivogolovko E.** The Influence of Data Quality on Clustering Outcomes. // Frontiers in Artificial Intelligence and Applications — IOS Press, 2012 — Vol. 249 — P. 95–105.