Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Кафедра высшей и прикладной математики

УТВЕРЖДАЮ

Заведующий кафедрой

_____ / С.Г. Мысливец

«17» июня 2021 г.

БАКАЛАВРСКАЯ РАБОТА

Направление 01.03.02 Прикладная математика и информатика

ИЗВЛЕЧЕНИЕ И ПРЕДСТАВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ ИЗ БИНАРНЫХ КОНТЕКСТОВ

Научный руководитель
кандидат физико-математических наук,
доцент _____ /Д.В. Семенова

Выпускник _____ /Е.А. Ленда

Красноярск 2021

РЕФЕРАТ

Выпускная квалификационная работа по теме «Извлечение и представление закономерностей из бинарных контекстов» содержит 49 страниц текста, 7 рисунков, 10 таблиц, 2 приложения и 36 использованных источников.

АНАЛИЗ ФОРМАЛЬНЫХ ПОНЯТИЙ, ФОРМАЛЬНЫЙ КОНТЕКСТ, ЗАМКНУТЫЕ МНОЖЕСТВА, АССОЦИАТИВНЫЕ ПРАВИЛА, КРИТЕРИИ ЗНАЧИМОСТИ, НЕИЗБЫТОЧНЫЙ БАЗИС, МИНИМАКСНЫЕ ПРАВИЛА.

Извлечение закономерностей позволяет выявлять часто встречающиеся сочетания элементов в базах данных в виде причинно-следственных связей и использовать обнаруженные связи для прогноза.

Средством описания причинно – следственных связей в базах данных, представленных матрицей «объект – признак», служат ассоциативные правила. Ассоциативных анализ оперирует довольно большим количеством разнообразных методов, применение которых ограничивается размерностью современных баз данных.

Цель работы — исследование моделей и алгоритмов реализации извлечения ассоциативных правил из бинарных контекстов.

В работе рассмотрены положения анализа формальных понятий. Изучены основные подходы к решению задачи поиска множества ассоциативных правил из заданного формального контекста. Разработаны программные модули, реализующие алгоритмы Apriori, Close и MClose для поиска ассоциативных правил из баз данных и проведены вычислительные эксперименты на реальных экономических контекстах, которые характеризуют уровень цифровизации рынков B2B,B2C и B2G.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Интеллектуальный анализ данных представляет собой процесс обработки информации с целью выявления неочевидных, объективных и потенциально полезных закономерностей. Данный термин был предложен Григорием Пятецким-Шапиро еще в 1989 году. Это время, когда вырос интерес к области исследования данных, а также к новым алгоритмам и методам решения конкретных задач в научно-технической, медицинской и экономической сферах. И по сей день технология обработки данных продолжает развиваться, чтобы идти в ногу с прогрессом, растущим потенциалом больших данных и доступной вычислительной мощностью.

За последние три десятилетия достижения в области информационных технологий позволили нам перейти от ручных и трудоемких методов к быстрому и автоматизированному анализу данных, который способен обрабатывать большие массивы информации. Чем обширнее собранные наборы данных, тем больше вероятность получения точной информации. Поставщики телекоммуникационных услуг, банки, страховые агенты, сотрудники продуктовых сетей и другие, используют интеллектуальный анализ данных для выявления взаимосвязей, начиная от оптимизации рабочего графика сотрудников, рекламных предложений и до оптимизации производственных процессов на предприятии, которые влияют на их бизнес-модели, операции и отношения к клиентам. Одной из важных дисциплин интеллектуального анализа данных является ассоциативный анализ. Он направлен на извлечение причинно-следственных связей между большими массивами данных, называемых ассоциативными правилами, которые показывают какие объекты встречаются вместе в базе и насколько часто.

Процесс обработки информации можно разделить на несколько этапов:

– Предварительная обработка данных. Этот шаг выполняется до применения алгоритмов ассоциативного анализа к интересующим данным. Предварительная обработка включает в себя очистку сырых данных, интеграцию, отбор и преобразование.
– Поиск ассоциативных правил. Осуществляется процесс непосредствен-

ного извлечения скрытых закономерностей в базе данных одним из наиболее подходящим методов.

– Анализ полученных данных. На этом этапе исследователь оценивает полученный результат в соответствии с требованиями пользователей и знаниями предметной области.

Основной проблемой поиска ассоциативных правил является огромное количество извлеченных правил. Многие из извлеченных правил считаются избыточными, поскольку они не предоставляют пользователю новой информации или они могут быть получены из других правил. Было предложено множество алгоритмов, которые уменьшают время генерации ассоциативных правил, а также предпринято много усилий по уменьшению размера извлеченного набора правил за счет перехода к «сжатым» формам представления множества ассоциативных правил. Особое интерес представляют алгоритмы, которые позволяют получать неизбыточный базис строгих ассоциативных правил.

Ассоциативные правила широко используются в различных областях, многие коммерческие предприятия накапливают огромное количество разнородной информации. Например, данные о клиентах продуктовых сетей. Компании заинтересованы в анализе данных, чтобы узнать о поведении своих покупателей. Полученная информация может быть использована для разработки новых бизнес-решений, таких как маркетинговые акции, управление запасами на предприятиях и взаимоотношениями с клиентами. Кроме того, различные онлайн-сервисы используют такие алгоритмы для оптимизации работы своих приложений, чтобы анализировать поведение посетителей, разрабатывать для них рекомендации. Различные подходы и методы ассоциативного анализа будут рассмотрены далее в данной работе.

Цели и задачи работы. Целью работы является исследование моделей и алгоритмов реализации извлечения ассоциативных правил из бинарных контекстов.

В рамках данной цели были поставлены и решены следующие задачи:

1. Исследовать существующие модели и алгоритмы поиска ассоциативных правил.

2. Программно реализовать алгоритмы извлечения ассоциативных правил для модели «поддержка – достоверность».
3. Провести обзор и анализ существующих критериев значимости.
4. Провести вычислительный эксперимент поиска ассоциативных правил на случайных и реальных контекстах.
5. Рассмотреть модели фильтрации полученных ассоциативных правил на основе изученных критериев значимости.

Краткое содержание работы.

В первой главе приведены основные идеи анализа формальный понятий. Рассматриваются основные подходы и алгоритмы нахождения множества ассоциативных правил.

Во второй главе представлен обзор критериев значимости для численной оценки ассоциативной связи, а также рассматриваются формальные требования к ним. Доказан ряд утверждений, которые позволяют упростить формулы для строгих ассоциативных правил.

В третьей главе описывается реализованный комплекс программных модулей для поиска ассоциативных правил и приведены результаты исследования уровня цифровизации рынков B2B (бизнес для бизнеса), B2C (бизнес для потребителя) и B2G (бизнес для государства).

ЗАКЛЮЧЕНИЕ

Результаты проведенной работы представлены ниже.

1. Рассмотрены основные подходы к поиску ассоциативных правил.
2. Разработаны программные модули, реализующие работу алгоритмов Apriori, Close, MClose.
3. Исследованы критерии значимости числовой оценки ассоциативных правил.
4. Проведены вычислительные эксперименты с использованием разработанных программ на реальных экономических контекстах, которые характеризуют рынки B2B, B2C, B2G.

Результаты бакалаврской работы докладывались на следующих конференциях и опубликованы в сборниках трудов этих конференций [13, 25]:

– XVI Международной конференции студентов, аспирантов и молодых ученых «Проспект Свободный — 2020» (диплом I степени).
– IEEE 14th International Conference on Application of Information and Communication Technologies AICT’2020.
– XVII Международной конференции студентов, аспирантов и молодых ученых «Проспект Свободный — 2021» (диплом I степени).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Биркгоф Г. Теория решеток / Г. Биркгоф — М.: Наука, 1984. — 568 с.
2. Быкова В. В. О неизбыточном представлении минимаксного базиса строгих ассоциативных правил / В. В. Быкова, А. В. Катаева // Прикладная дискретная математика. — 2017. — №36. — С. 113–126.
3. Быкова В. В. Алгоритм построения неизбыточного минимаксного базиса строгих ассоциативных правил / В. В. Быкова, А. В. Катаева // Прикладная дискретная математика. Приложение. — 2017. — №10. — С. 154–157.
4. Быкова В. В. Сжатое представление строгих ассоциативных правил в анализе данных / В. В. Быкова, А. В. Катаева // Программные продукты и системы. — 2017. — №2 (30). — С. 187–192.
5. Бююль А. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цефель. — СПб.: ООО ДиаСофтЮП, 2005. — 608 с.
6. Вапник В .Н. Теория распознавания образов: статистические проблемы обучения / В .Н. Вапник, А. Я. Червоненкис. — М.: Наука, 1974. — 416 с.
7. Витяев Е. Е., Демин А. В., Пономарев Д. К. Вероятностное обобщение формальных понятий/ Е. Е. Витяев, А. В. Демин, Д. К. Пономарев, // Программирование. — 2012. — Т. 38. — № 5. — С. 18-34.
8. Городецкий В.И., Самойлов В.В. Ассоциативный и причинный анализ и ассоциативные байесовские сети // Труды СПИИРАН. 2009. № 9. С. 13-65.
9. Гуров С. И. Булевы алгебры, упорядоченные множества, решетки: определения, свойства, примеры / С. И. Гуров. — М.: Либроком, 2013. — 221 с.
10. Игнатов Д. И. Бикластеризация объектно-признаковых данных на основе решеток замкнутых множеств Д. И. Игнатов, С О. Кузнецов // Труды 12-й национальной конф. по искусственному интеллекту с междунар. участием. — 2010. — С. 175–182.
11. Катаева А. В. Минимизация базиса строгих ассоциативных правил на основе замкнутых множеств / А. В. Катаева // Материалы республиканской

научно-практической конф. «Статистика и ее применения». — Ташкент, 2017. — С. 77–83.

12. Кузнецов С. О. Методы теории решеток и анализа формальных понятий в машинном обучении / С. О. Кузнецов // Новости искусственного интеллекта. — 2004. — №3. — С. 19–31.
13. Ленда, Е.А. Методы и алгоритмы построения неизбыточного минимаксного базиса ассоциативных правил / Е. А. Ленда, Д.В. Семенова // Проспект Свободный –2020: материалы XVI Международной конференции студентов, аспирантов и молодых ученых. – Красноярск, 2020.–С. 828-830.
14. Pasquier N., Bastide Y., Taouil R., Lakhal L. Efficient mining of association rules using closed itemset lattices // Information systems. – 1999. – Vol. 24. – No. 1. – P. 25-46.
15. Stumme G., Taouil R., Bastide Y., Pasquier N., Lakhal L. Computing iceberg concept lattices with TITANIC // Data knowledge engineering. – 2002. – Vol. 42. – No. 2. – P. 189-222.
16. Agrawal R. Database mining: A performance perspective / R. Agrawal, T. Imielinski, A. Swami // Special Issue on Learning and Discovery in Knowledge-Based Databases / Ed. by N. Cercone, M. Tsuchiya. — Washington, U.S.A.: Institute of Electrical and Electronics Engineers, 1993. — Vol. 6. — No. 5. — P. 914–925.
17. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases // Proceedings of ACM SIGMOD International Conf. on Management of Data. In P. Buneman, S. Jajodia (eds.). 1993. pp. 207-216.
18. Agrawal R. Fast algorithms for mining association rules. R. Agrawal, R. Srikant // Proceedings 20 th Int. Conf. Very Large Data Bases, VLDB. — Morgan Kaufmann, 1994. — P. 487–499.
19. Armstrong W. W. Dependency structure of data bases relationships / W. W. Armstrong// Proceedings IFIP Congress. Geneva. — 1974. — P. 580–583.
20. Aze J. Extraction de pepites de connaissances dans les donnees: Une nouvelle approche et une etude de sensibilite au bruit. In Mesures de Qualite pour

la fouille de donnees. / J. Aze and Y. Kodratoff // Revue des Nouvelles Technologies de l'Information, Brest, France, 2004.
21. Clark P., Boswell R. Rule induction with cn2: some recent improvements // Proceedings of the European Working Session on Learning EWSL-91, Porto, Portugal. 1991. pp. 151-163.
22. Clark P, Brin S., Motwani R., Ullman J., Tsur S. Dynamic itemset counting and implication rules for market basket data // Proceedings of ACM-SIGMOD International Conference on Management of Data, Montreal, Canada. 1997. pp. 255-264.
23. Ganter B. Formal Concept Analysis: Foundations and Applications / B. Ganter, G. Stumme, R. Wille — Berlin Heidelberg: Springer, 2005. — 315 p.
24. Geng L., Hamilton H. Interestingness measures for data mining: A survey // ACM Computing Surveys (CSUR). — 2006. — Vol. 38. — No. 3. — P. 9-41.
25. Goldenok E. E. Association Analysis of Digital Transformation Processes of the B2G Trade Services Market Segment in Russia / E. E. Goldenok, E. A. Lenda, D. V. Semenova, A. K. Yakobson // 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT). — 2020. — P. 514–517.
26. Han J. Re-examination of interestingness measures in pattern mining: a unified framework. / T. Wu, Y. Chen, J. Han — Data Mining and Knowledge Discovery. — 2009. — P. 371-397.
27. Lenca P., Vaillant B., Meyer P., Lallich S. Association Rule Interestingness Measures // Experimental and Theoretical Studies. Quality Measures in Data Mining. 2007. Vol. 43. pp. 51-76.
28. Mosteller F. Association and estimation in contingency tables // Journal of American Statistical Association. 1968. No. 63 (321). pp. 1-26
29. Pasquier N. Efficient mining of association rules using closed itemset lattices / N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal // Information systems. — 1999. — Vol. 24. — No. 1. — P. 25-46.
30. Piatetsky–Shapiro G. Discovery, analysis and presentation of strong rules / Knowledge Discovery in Databases. AAAI Press/MIT Press. 1991. P. 229–248

31. Sahar S., Mansour Y. An empirical evaluation of objective interestingness criteria // Proceeedings of SPIE Conference on Data Mining and Knowledge Discovery, Orlando, FL. 1999. pp. 63-74.
32. Sebag M., Schoenauer M. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases // Proc. of the European Knowledge Acquisition Workshop EKAW'88. 1988. pp. 28.1-28.20.
33. Tan P. Interestingness Measures for Association Patterns: A Perspective. / P. Tan and V. Kumar // In Proceedings of the Workshop on Postprocessing in Machine Learning and Data Mining, New York, USA, 2000.
34. Tan P. Selecting the right objective measure for association analysis / P. Tan, V. Kumar and J. Srivastava // Information Systems, — Vol. 29(4), 2004. —P. 293–313.
35. Yule G.U. On the methods of measuring association between two attributes // J. R. Stat. Soc. 75. 1912. pp. 579-642. 196
36. Zaki M. J. Efficient algorithms for mining closed itemsets and their lattice structure / M. J. Zaki, C. J. Hsiao // IEEE Transactions on Knowledge Data Engineering. – 2005. — No. 4. — P. 462-478.

Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Кафедра высшей и прикладной математики

УТВЕРЖДАЮ

Заведующий кафедрой

_____ / С.Г. Мысливец

«17» июня 2021 г.

БАКАЛАВРСКАЯ РАБОТА

Направление 01.03.02 Прикладная математика и информатика

ИЗВЛЕЧЕНИЕ И ПРЕДСТАВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ ИЗ БИНАРНЫХ КОНТЕКСТОВ

Научный руководитель
кандидат физико-математических наук,
доцент _____ /Д.В. Семенова
17 июня 2021

Выпускник _____ /Е.А. Ленда
17 июня 2021

Красноярск 2021