Система анализа данных коллаборативных платформ CrowDM

Д. И. Игнатов[1], А. Ю. Каминская[2], А. А. Беззубцева[3], К. Н. Блинкин[4]

[1]dignatov@hse.ru, [2]skam90@gmail.com, [3]nstbezz@gmail.com
[4]xkonstantinx@gmail.com

НИУ ВШЭ, Россия, 101000, г. Москва, ул. Мясницкая, д. 20

Аннотация. В работе описывается система анализа данных коллаборативной платформы компании Witology. Проект находится в состоянии разработки, поэтому в статье отражены в основном методологические аспекты и результаты первых экспериментов. В основу системы положен ряд моделей и методов современного анализа объектно-признаковых и неструктурированных данных (текстов), таких как Анализ Формальных Понятий, мультимодальная кластеризация, поиск ассоциативных правил и извлечение ключевых словосочетаний и слов из текстов.

Ключевые слова: коллаборативные и краудсорсинговые платформы, разработка данных (Data Mining), анализ формальных понятий, мультимодальная кластеризация.

Введение

Успехи современной индустрии коллаборативных технологий ознаменовались появлением ряда новых платформ для проведения распределенных мозговых штурмов или осуществления так называемой общественной экспертизы, например, на Российском рынке такие продукты выпускают компании Witology [1] и Wikivote [2]. И, хотя до технологического прорыва еще далеко, несколько крупных проектов уже

Игнатов Д.И., Яворский Р.Э. (ред.): Анализ Изображений Сетей и Текстов, Екатеринбург, 16-18 марта, 2012

успешно завершены. Среди них «Сбербанк-21», анализ форумов Агентства Стратегических Инициатив и др. Массивы данных нового типа систем, ядро которых составляют так называемые социосемантические сети, требуют новых подходов к анализу данных. В рамках данной статьи мы предлагаем новую методологическую базу для анализа данных коллаборативных систем, опирающуюся на современные модели и методы разработки данных (Data Mining) и искусственного интеллекта.

Как правило, в рамках одного проекта пользователи таких краудсорсинговых платформ [3] решают некую общую задачу, выдвигают идеи, оценивают идеи друг друга как эксперты, а в итоге по результатам обсуждений и рейтингования определяются лучшие идеи и люди – генераторы идей. Для более глубокого понимания поведения пользователей, выработки адекватных критериев оценки, анализа динамики и статистики в ходе развития проекта необходимы особые средства. Традиционные методы кластеризации, поиска сообществ и анализа текстов нуждаются в адаптации, а иногда и в полной переработке, требуют изобретательности для их результативного применения, т.е. получения действительно полезных и нетривиальных результатов. Мы кратко описываем модели данных, используемых в проекте, в терминах Анализа Формальных Понятий (АФП) [4]. Также мы приводим описание системы анализа данных CrowDM (Crowd Data Mining), ее архитектуру и методы, лежащие в основе ключевых этапов анализа данных.

Математические модели и методы

На начальном этапе анализа данных коллаборативной платформы были выявлены два типа данных такой платформы, напрямую соответствующие двум составляющим социосемантической сети: данные без использования ключевых слов (связи, оценки, действия пользователей) и данные с ключевыми словами (наполнение всего создаваемого контента на платформе).

Для анализа данных без ключевых слов предлагается применять методы анализа социальных сетей (Social Network Analysis), кластеризации (а также би- и трикластеризации [5, 6, 7, 8], спектральной кластеризации), анализ формальных понятий (решетки понятий, импликации, ассоциативные правила) и его расширения для случая мультимодальных данных, например, триадических [9]; рекомендательные системы [10, 11,12] и статистические методы анализа (анализ распределений и средних значений).

Для методов анализа текстовых данных с использованием ключевых слов, основным является этап выделения ключевых слов и словосочетаний. Это направление компьютерной лингвистики заслуживает от-

дельного рассмотрения, поэтому в данной статье мы остановимся на некоторых методах анализа данных без использования ключевых слов. На схеме анализа (см. рис. 2) синим цветом выделены методы, описанные в данной статье.

Главными действующими лицами в краудсорсинговых проектах, а значит и в коллаборативных платформах, созданных для этих проектов, являются пользователи платформы, они же участники проекта. Будем рассматривать их в качестве *объектов* для анализа. Вместе с тем, каждый объект может обладать (или не обладать) определенным набором *признаков*. В качестве признаков пользователей коллаборативной платформы могут выступать темы, в обсуждении которых пользователь принимал участие, идеи, которые он выдвигал или за которые голосовал, и даже другие пользователи. Основным инструментом для анализа данных объектно-признаковой природы является анализ формальных понятий (АФП). Дадим формальные определения.

Контекстом в АФП называют тройку $\mathbb{K} = (G, M, I)$, где G — *множество объектов*, M — *множество признаков*, а отношение $I \subseteq G \times M$ говорит о том, какие объекты какими признаками обладают. Для произвольных $A \subseteq G$ и $B \subseteq M$ определены *операторы Галуа*:

$$A' = \{m \in M \mid \forall g \in A\ (g\ I\ m)\};$$
$$B' = \{g \in G \mid \forall m \in B\ (g\ I\ m)\}.$$

Оператор '' (двукратное применение оператора ') является *оператором замыкания*: он идемпотентен ($A'''' = A''$), монотонен ($A \subseteq B$ влечет $A'' \subseteq B''$) и экстенсивен ($A \subseteq A''$). Множество объектов $A \subseteq G$, такое, что $A'' = A$, называется *замкнутым*. Аналогично для замкнутых множеств признаков — подмножеств множества M. Пара множеств (A, B), таких, что $A \subseteq G$, $B \subseteq M$, $A' = B$ и $B' = A$, называется *формальным понятием* контекста $\mathbb{K}$. Множества A и B замкнуты и называются *объемом* и *содержанием* формального понятия (A, B) соответственно. Для множества объектов A множество их общих признаков A' служит описанием сходства объектов из множества A, а замкнутое множество A'' является кластером сходных объектов (с множеством общих признаков A'). Отношение “быть более общим понятием” задается следующим образом: $(A, B) \geq (C, D)$ тогда и только тогда, когда $A \supseteq C$. Понятия формального контекста $\mathbb{K} = (G, M, I)$, упорядоченные по вложению объемов образуют решетку $\underline{\mathfrak{B}}(G, M, I)$, называемую *решеткой понятий*. Для визуализации решеток понятий используют т.н. диаграммы Хассе, т.е. граф покрытия отношения “быть более общим понятием”.

Так как в худшем случае (булева решетка понятий) количество понятий равно $2^{\min\{|G|,|M|\}}$, то для больших формальных контекстов разумно применять АФП, если данные разрежены. Так же можно использовать различные способы сокращения количества формальных понятий, такие как отбор понятий по индексу устойчивости или размеру объема. Альтернативным подходом является ослабление определения формального понятия, как максимального прямоугольника в объектно-признаковой матрице все элементы которого принадлежат отношению инцидентности. Одним из таких ослаблений является определение объектно-признакового бикластера [2,3].

Если $(g, m)\in I$, то (m', g') называется *объектно-признаковым бикластером* с плотностью $\rho(m',g')=|I\cap(m'\times g')|/(|m'|\cdot|g'|)$.

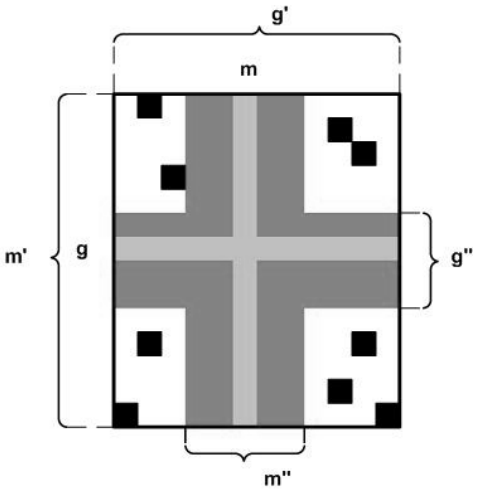

рис. 1. оп-бикластер

Приведем основные свойства оп-бикластеров:

1. для любого бикластера $(A,B)\in 2^G\times 2^M$ выполняется $0\le\rho(A,B)\le 1$.

2. оп-бикластер (m',g') является формальным понятием тогда и только тогда, когда $\rho=1$.

3. Если (m', g') – бикластер, то $(g'', g')\le(m', m'')$.

Пусть $(A,B)\in 2^G\times 2^M$ будет бикластером и ρ_{min} неотрицательное действительное число такое, что $0\le\rho_{min}\le 1$, тогда (A, B) называется *плотным*, если он удовлетворяет ограничению $\rho(A,B)\ge\rho_{min}$.

Из вышеописанного следует, что оп-бикластеры отличаются от формальных понятий тем, что в них не обязательно наблюдается единичная плотность. Графически это означает, что не обязательно все «ячейки» на пересечении объектов и признаков бикластера должны быть заполнены (см. рис. 1).

Помимо построения решеток понятий и их визуализации с помощью диаграмм Хассе используются импликации и ассоциативные правила для выявления признаковых зависимостей в данных. Далее на основе полученных результатов, можно формировать рекомендации, например, предлагать пользователям наиболее интересные для них обсуждения. Кроме того, можно произвести структурный анализ сети и применить методы кластеризации для поиска сообществ, а также статистические методы для частотного анализа различной активности пользователей.

Почти все вышеперечисленные методы можно применять и к данным с использованием ключевых слов, отличие состоит лишь в том, что в качестве признаков будут выступать ключевые слова, например, употребляемые конкретным пользователем или группой пользователей.

Схема анализа

Схема анализа данных системы CrowDM, создаваемой в данный момент проектно-учебной группой НИУ ВШЭ, представлена на рисунке 2. Ранее упоминалось, что после выгрузки данных из базы, мы получаем формальные контексты и коллекции текстов. Последние в свою очередь тоже преобразуются в формальные контексты после выделения ключевых слов. Далее анализируются полученные контексты.

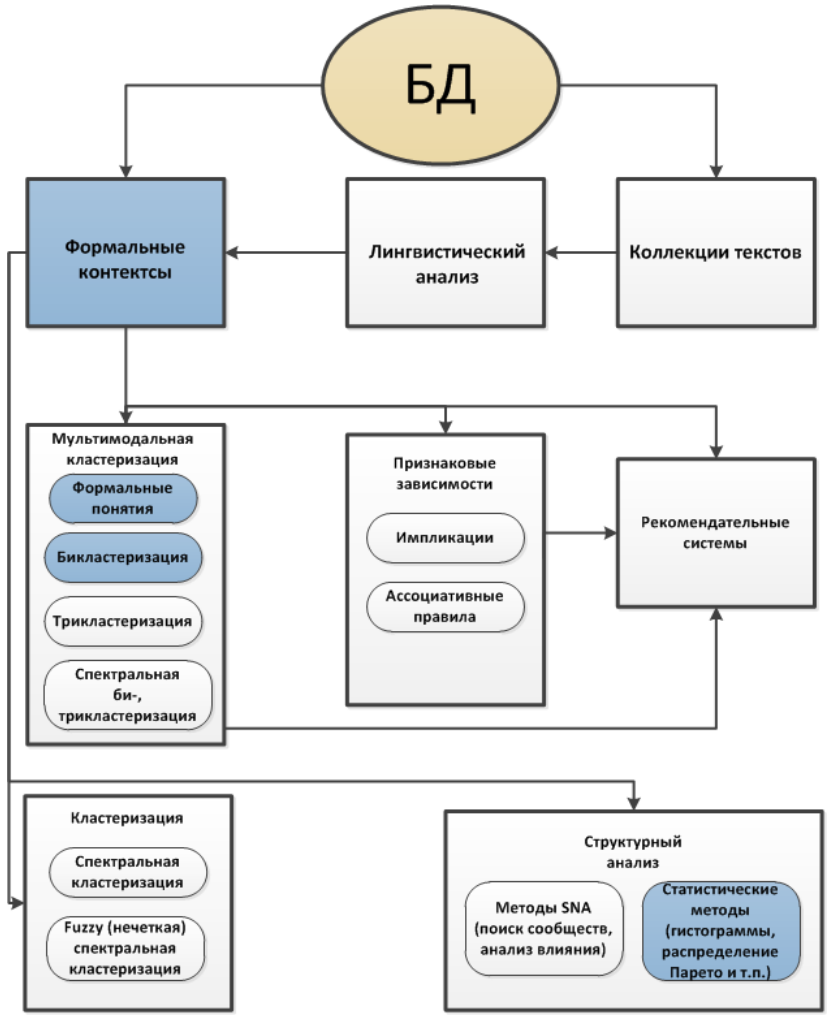

рис. 2. Схема анализа данных коллаборативных платформ в системе CrowDM

Результаты экспериментов

Для проведения первых двух экспериментов были отобраны формальные контексты, в которых в качестве объектов выступают пользователи платформы, а в качестве признаков – идеи, которые они предлагали в рамках одной из пяти тем проекта («Сбербанк и частный клиент»). Из всех идей были также отобраны лишь те, которые дошли по-

чти до самого конца проекта. Считается, что объект «пользователь» обладает признаком «идея», если данный пользователь внес любой вклад в обсуждение идеи: является автором идеи, комментировал идею, оставил комментарий в ветке этой идеи, выставил оценку этой идее или комментариям к ней. Таким образом, найденные формальные понятия вида (U,I), где U – множество пользователей, I – множество идей, соответствуют так называемым эпистемическим сообществам (проще говоря, сообществам по интересам) из множества людей U, которые интересуются множествами идей I.

На рисунке 3 представлена диаграмма полученной решетки понятий.

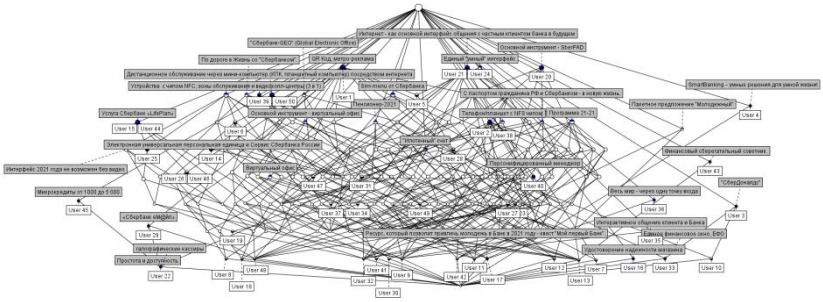

рис. 3. Диаграмма решетки формальных понятий для контекста пользователи-идеи.

Каждому узлу диаграммы решетки соответствует одно формальное понятие (в данной решетке всего 198 понятий). Также каждый узел помечен множеством объектов и признаков, если этот узел является первым, где встречается данный объект (при движении снизу вверх по диаграмме) или признак (при движении сверху вниз) соответственно. Очевидно, что полученная диаграмма решетки является достаточно громоздкой для анализа по ее статическому изображению. Обычно в таких случаях для визуализации используют порядковые фильтры (верхняя часть решетки) или диаграммы множества устойчивых понятий. Мы в свою очередь демонстрируем отдельный фрагмент решетки (см. рис. 4), таким образом, объясняя способ ее «чтения».

Эксперименты были проведены в программе Concept Explorer, разработанной специально для применения алгоритмов АФП к объектно-признаковым данным. Выделив любой узел решетки, можно увидеть объекты и признаки, соответствующие понятию в этом узле. Объекты «накапливаются» снизу (в данном примере множество объектов состоит из User45 и User22), признаки – сверху (у нас один признак – «Микро-

кредиты от 1000 до 5000»). Это означает, что пользователи User45 и User22 вместе участвовали в обсуждении идеи с указанным именем и больше ни один из пользователей участия в обсуждении не принимал.

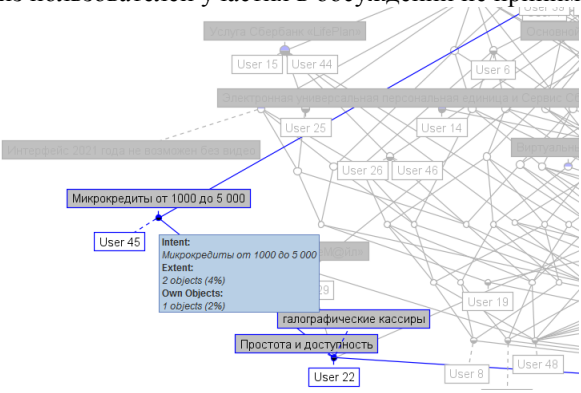

рис. 4. Фрагмент диаграммы решетки понятий

Ниже представлены результаты применения алгоритмов бикластеризации на тех же самых данных.

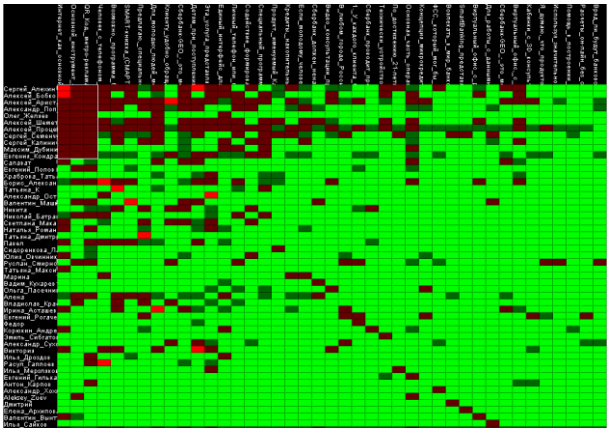

рис. 5. Результат работы алгоритма бикластеризации BiMax

Поясним рисунок 5. Эксперименты проведены в системе анализа данных генной экспрессии BicAT. Строки соответствуют пользователям, столбцы – идеям в рамках указанной темы, в обсуждении которых пользователи принимали участие. Цвет ячейки на пересечении соответствующей строки и столбца соответствует интенсивности вклада конкретного пользователя в данную проблему. Под вкладом пользователя

понимается взвешенная сумма числа его комментариев к этой идее, количества оценок, при этом учитывается, является ли данный человек автором этой идеи, или нет. Самые светлые ячейки соответствуют нулевому вкладу, самые яркие (см. левую верхнюю ячейку на рис.6) – максимальному вкладу. После дискретизации данных (0 соответствовал нулевому вкладу, 1 – ненулевому) к ним был применен алгоритм бикластеризации BiMax, который нашел несколько бикластеров (см. пример на рисунке 6). Поскольку одной из задач проведения краудсорсинговых проектов является поиск людей со схожими идеями, представленный бикластер из 11 пользователей наиболее интересен, в то время как остальные найденные бискластеры содержали в среднем по 4-5 пользователей (с ограничением на количество идей в бикластере строго больше двух).

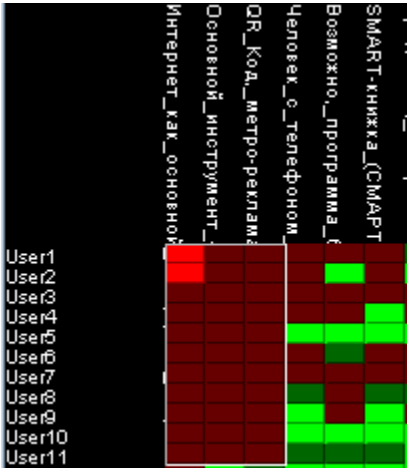

рис. 6. Бикластер с большим числом пользователей

Далее, чтобы более полно увидеть картину оценивания в проекте, было построено несколько видов графиков распределения оценок. Одним из примеров является график на рисунке 6, который отображает кумулятивное число пользователей, выставивших больше определенного количества оценок за весь проект.

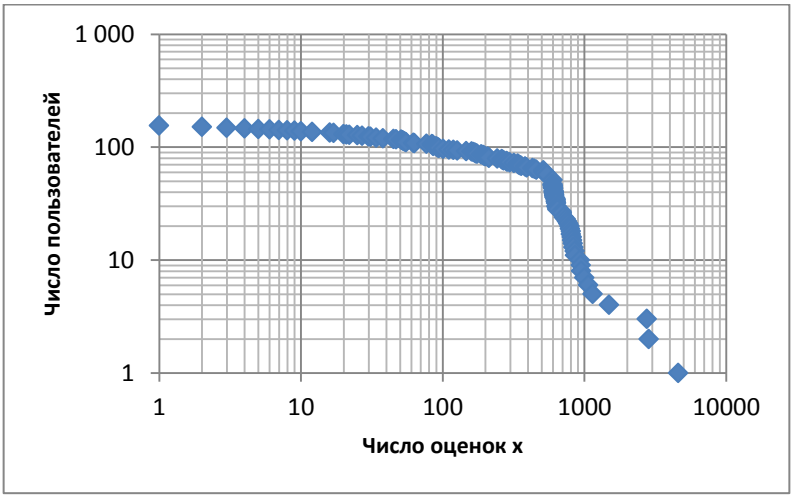

рис. 7. Распределение числа оценок

По оси абсцисс отложено количество оценок, оставленных пользователем. По оси ординат – число пользователей, которые выставили больше соответствующего числа оценок. Например, больше 5000 оценок поставил один пользователь (крайняя правая точка на оси абсцисс), а больше 4000 – уже упомянутый пользователь и еще один участник. Всего участников, поставивших хотя бы одну оценку, 167. Множество точек явно разделяется на две части: пологая длинная линия (от x=0 до 544 включительно) и более крутой хвост. Тот факт, что в логарифмических шкалах обе части выглядят похожими на прямые, указывает на то, что обе части, возможно, распределены по Парето.

Целесообразно искать отдельные функции распределения для основной и хвостовой части выборки, потому как если проверить всю выборку на соответствие, например, Парето-распределению, нулевая гипотеза о соответствии отвергается на близком к нулю уровне значимости.

Заключение

Результаты первых экспериментов позволяют утверждать, что разрабатываемая методология окажется полезной для анализа данных коллаборативных систем и систем совместного пользования ресурсами.

Среди направлений дальнейшей работы наиболее приоритетными являются использование текстовой информации генерируемой пользователем и применение методов мультимодальной кластеризации, а также создание рекомендательных сервисов на их основе.

Благодарности

Работа выполнена в рамках проектно-учебной группы НИУ ВШЭ “Алгоритмы интеллектуального анализа данных (Data Mining) для Интернет-форумов обсуждения инновационных проектов”.

Список источников

1. http://witology.com/

2. http://www.wikivote.ru/

3. Jeff Howe. The Rise of Crowdsourcing. Wired, 2006.

4. Ganter, B., Wille, R. Formal Concept Analysis. Springer, Heidelberg, 1999.

5. Игнатов Д.И., Кузнецов С.О. Бикластеризация объектно-признаковых данных на основе решеток замкнутых множеств// Труды 12-й национальной конференции по искусственному интеллекту, М., Физматлит, Т. 1., С.175-182, 2010.

6. Игнатов Д.И., Каминская А.Ю., Кузнецов С.О., Магизов Р. А. Метод бикластеризации на основе объектных и признаковых замыканий// Интеллектуализация обработки информации: 8-я международная конференция. Республика Кипр, г. Пафос, 17-24 октября 2010 г.: Сборник докладов.– М.: МАКС Пресс, 2010. – С. 140 – 143.

7. Игнатов Д.И., Магизов Р.А. Анализ тримодальных данных на примере Интернет-сервисов социальных закладок// Социологические методы в современной исследовательской практике: Сборник статей, посвященный памяти первого декана факультета социологии НИУ ВШЭ А.О. Крыштановского / Отв. ред. и вступит. ст. О.А. Оберемко; НИУ ВШЭ, ИС РАН, РОС. М.: НИУ ВШЭ, 2011.

8. Игнатов Д. И., Кузнецов С. О., Пульманс Й. Разработка данных систем совместного пользования ресурсами: от трипонятий к трикластерам //Математические методы распознавания образов: 15-я Всероссийская конференция. г. Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. — М.: МАКС Пресс, 2011. — 618 с. (ISBN 978-5-317-03787-1)

9. Robert Jäschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, Gerd Stumme: TRIAS - An Algorithm for Mining Iceberg Tri-Lattices. ICDM 2006: 907-911

10. Игнатов Д.И., Кузнецов С.О. Методы разработки данных (Data Mining) для рекомендательной системы Интернет-рекламы // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2008, 28 сентября – 3 октября 2008 г., г. Дубна, Россия): Труды конференции. Т.2. – М.: Ленанд, 2008. – 392 с.

11. D.I. Ignatov, S.O. Kuznetsov. Concept-based Recommendations for Internet Advertisement// In proceedings of The Sixth International Conference Concept Lattices and Their Applications (CLA'08), Radim Belohlavek, Sergei O. Kuznetsov (Eds.): CLA 2008, pp. 157–166 ISBN 978–80–244–2111–7, Palacky University, Olomouc, 2008.

12. Dmitry I. Ignatov, Sergei O. Kuznetsov, Ruslan A. Magizov and Leonid E. Zhukov. From Triconcepts to Triclusters// In proceedings of 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Kuznetsov et al. (Eds.): RSFDGrC 2011, LNCS/LNAI Volume 6743/2011, Springer-Verlag Berlin Heidelberg, 257-264, 2011.