Анализ Больших Данных для сегментирования востребованности услуг малого бизнеса по направлениям деятельности в регионе

В.М. Рамзаев[а], И.Н. Хаймович[а,б], В.Г. Чумак[а]

[а] *Международный институт рынка, 443030 ,ул. Г.С. Аксакова, 21, Самара, Россия*
[б] *Самарский национальный исследовательский университет им. академика С.П. Королева, 443086, Московское шоссе, 34, Самара, Россия*

Аннотация

В статье предложен инструмент для повышения эффективности использования бюджетных средств в регионе в области малого бизнеса. Это является важнейшей задачей в современных экономических условиях, в основе решения которой лежит возможность принятия оптимальных управленческих решений. Предложенный способ регулирования на основе анализа социальных сетей с использованием технологии BIG DATA может быть эффективен при управлении различными инновационными процессами развития экономики региона, для которых характерны многообразие форм и широкий спектр составляющих и факторов, а также свойственна динамика развития и активная трансформации жизнедеятельности. Использование современных программно-аппаратных средств из технологии BIG DATA позволяет производить оценку и визуализацию изменений фактически в режиме реального времени.

Ключевые слова: конкурентоспособность; управление территориями; интенсивные данные; математические модели; технология BIG DATA

1. Введение

В современных социально-экономических условиях актуальной задачей является государственное регулирование субъектов рыночной экономики, среди которых одним из важнейших в регионе выступает малый бизнес (МБ). Зарубежный опыт показывает, что без этого сектора народного хозяйства невозможно развитие экономики, поскольку от него зависят темпы экономического роста, структура и качество до 40 - 50% валового национального продукта.

2. Объект исследования

Структура малых и средних предприятий по видам экономической деятельности не однородна. Как видно из рисунка 1, наибольшее число предприятий занимается торговлей, ремонтом автотраспортных средств, мотоциклов, бытовых изделий и предметов личного пользования, что объясняется более низкими входными барьерами в эти сферы деятельности.

Можно выделить следующие особенности управления развитием МБ. Во-первых, необходимо отметить широкий спектр услуг, оказываемых субъектами МБ, а также огромный ассортимент реализуемых ими товаров. Во-вторых, МБ отличается существенно большей мобильностью по сравнению с крупным. Под мобильностью мы понимаем постоянное изменение конъюнктуры рынка, закрытие старых и появление новых хозяйствующих субъектов, что объясняется высокой вариативностью вкусов и предпочтений потребителей товаров и услуг субъектов МБ, т.е. идет достаточно активный процесс замещения одних видов деятельности другими, детерминированный изменением платежеспособного спроса, что особенно актуально в современных условиях импортозамещения. Так, согласно статистике, до 85% новых субъектов МБ закрывается в первый год своего существования. Из 100 зарегистрированных малых предприятий к четвертому году прекращают деятельность 94 малых предприятия.

В связи с этим применение традиционных методов государственного управления, опирающихся на данные месячной, квартальной и годовой статистики, не приносит ожидаемого результата и не позволяет выявить тренды на развитие или сворачивание тех или иных видов деятельности, поэтому зачастую принятие решений о финансовой поддержке и выделении средств на те или иные проекты существенно отстает от потребностей, а в ряде случаев и противоречит изменившейся реальной рыночной ситуации к моменту начала финансирования.

Например, в настоящее время в Самарском регионе господдержка малому и среднему предпринимательству реализуется в рамках осуществления Государственной программы «Развитие предпринимательства, торговли и туризма в Самарской области» на 2014 - 2019 годы, утвержденного постановлением Правительства Самарской области от 29.11.2013 №699. Помощь бизнесменам Самарской области оказывается по разным направлениям и заключается в информационно-консультационных услугах, обучении, финансовой помощи, содействии в реализации товаров и услуг.

Вместе с тем следует отметить, что, несмотря на целый ряд мер, используемых властью в регионе для управления развитием МБ, на сегодняшний момент не разработаны эффективные методы выбора приоритетных направлений развития МБ, позволяющие наиболее целесообразно направлять бюджетные средства на развитие и поддержку предпринимателей [1-4]. Рынок малого и среднего предпринимательства – достаточно динамично меняющаяся среда. Это необходимо принимать во внимание при среднесрочном и долгосрочном планировании и должно быть взято за основу органами власти региона при поддержке и стимулировании развития наиболее востребованных направлений деятельности МБ и мониторинга эффективности использования бюджетных средств на программы в данной отрасли предпринимательства в условиях изменяющейся конъюнктуры рынка.

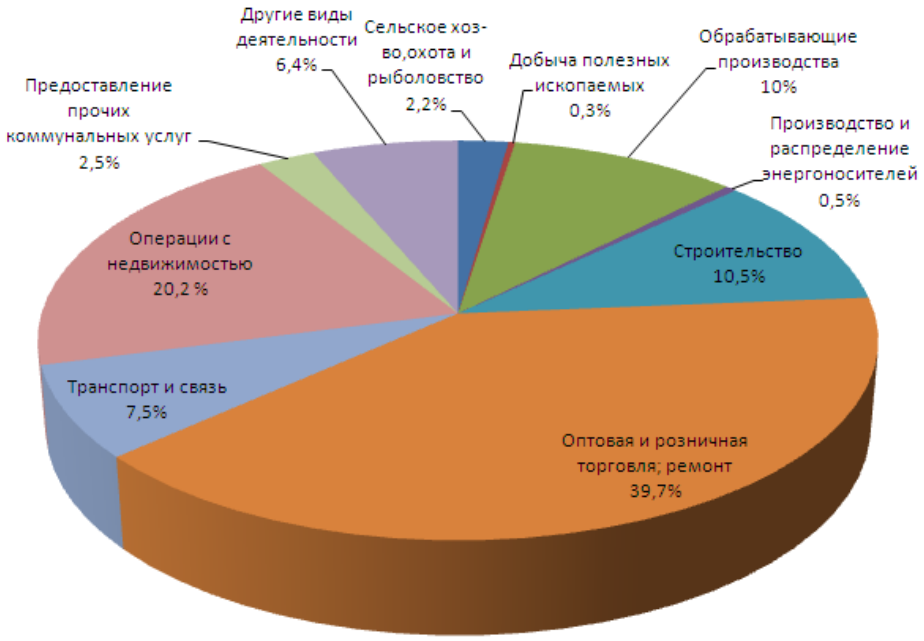

Рис. 1. Структура малых и средних предприятий по видам экономической деятельности на конец 2014 г., %.

3. Методы использования интеллектуального анализа данных при определении сегментов малого бизнеса в регионе

Эту задачу возможно решить с использованием современных информационных технологий [5,6,7], к которым относится технология BIG DATA, непосредственно связанная с интеллектуальным анализом данных [8,10]. Вместе с тем применение современных технологий BIG DATA позволяет выделить зоны – территории наиболее активного потребления и востребованности тех или иных видов товаров и услуг в режиме реального времени на рынке.

Для управления развитием малого и среднего бизнеса в регионе на основе BIG DATA была разработана специальная методология [9], состоящая из следующих этапов: определение роли и места малого бизнеса в регионе; определение основных видов и услуг, предлагаемых малым бизнесом в регионе; создание портрета потребителя, пользующегося услугами малого бизнеса; создание информационной модели потребителя малого бизнеса в регионе; формирование зон малого бизнеса в регионе; разработка рекомендаций по принятию управленческих решений.

Если роль и место малого и среднего бизнеса в регионе, основные виды и услуги, предлагаемые предпринимателями в регионе, были проанализированы, то для создания портрета и информационной модели потребителя необходимо использовать технологии BIG DATA. Метод использования интеллектуального анализа данных состоит в следующем:

1. формирование набора больших данных в hadoop из twitter по фильтру «Самарская область», выявляющему количество обращений;
2. разделение сформированного набора по различным фильтрам, связанным с базовыми факторами малого бизнеса;
3. проведение мониторинга потокового анализа неструктурированной информации по фильтрам;
4. принятие оперативных мероприятий в случаях устойчивых «всплесков» по количеству обращений;
5. разработка программы на языке Scala для работы с фильтрацией в области Больших Данных;
6. отладка и тестирование программы с набором практических данных;
7. анализ результатов вычисления.

Для получения данных используется социальная сеть «twitter», так как это «открытый» продукт, его применение не требует дополнительных инвестиций, а 50% пользователей Интернет имеют профили в данной программе. Twitter является второй по популярности сетью среди пользователей во всем мире, уступая лишь Facebook. Однако в отличие от Facebook, который не предоставляет открытый доступ к своим данным, Twitter такой доступ предоставляет, отсутствуют ограничения на доступ к наборам данных сервера. Пользователи данной социальной сети обмениваются в основном текстовой информацией, что является несомненным плюсом при обработке. Twitter не является предметной сетью и наиболее широко отражает общественное мнение по многим интересующим вопросам, поэтому для формирования зон малого бизнеса в регионе обработка данных из этой социальной сети была оптимальной.

Для работы с BIG DATA в социальных сетях используют методы сбора, обработки и анализа данных. Сбор данных осуществляется в режиме реального времени, в пределах определенной геолокации, либо в пределах всей сети, по определенным шаблонам. Информация, представляющая интерес для анализа в области МБ, это: локация, дата и время, контент, «автор» контента (пользователь), связи между пользователями. Сбор данных в социальных сетях можно осуществлять с помощью следующих инструментов: Apache Hadoop, Biglnsights (IBM), Cloudera, Hortonworks, Storm. Для выполнения исследования в области МБ был выбран Hortonworks. Для работы применялся Twitter Application (apps.twitter.com), в котором определялись и уточнялись ключевые параметры: API key, API secret, Access token, Access token secret.

Для сбора данных с использованием Hortonworks, Twitter App использовался конфигурационный файл сервиса flume в виртуальной машине Hortonworks Sandbox. После установки виртуальной машины Hortonworks_ Sandbox версии 2.3 и настройки сервиса flume система готова к загрузке данных из twitter. Для просмотра и загрузки скаченных файлов переходим в папку HDFS, где осуществляем обработку данных. Вид файловой структуры HDFS в виртуальной машине Hortonworks при решении задачи в сфере МБ показан на рисунке 2.

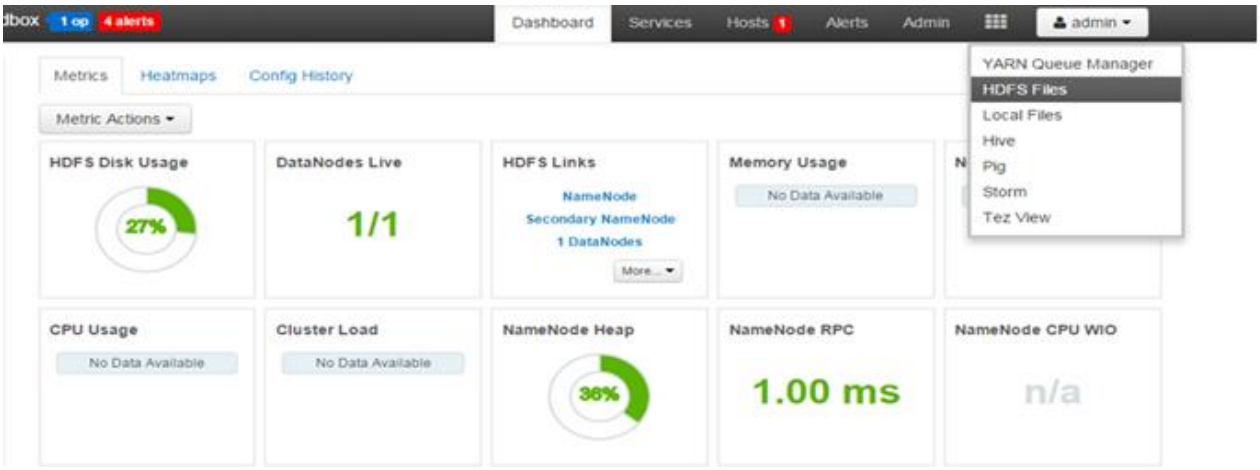

Рис. 2. Визуализация HDFS в Hortonworks при загрузке файлов для решения задачи в области МБ.

Собранные данные необходимо структурировать (т.е. обработать) в соответствии с парадигмой MapReduce. MapReduce — это фреймворк для выполнения распределенных задач с использованием большого количества компьютеров, образующих кластер.

Использование MapReduce позволило структурировать поток данных из социальных сетей по критериям: шрифты, размер текста, цвет, ссылка на профайл пользователя, локация, время и прочее.

Для определения портрета потребителя для МБ в нашем исследовании нужны данные следующих типов: размещение, текст, язык и время. Для того, чтобы извлечь лишь эту информацию можно использовать технологию MapReduce, встроенную в инструменте Hortonworks Sandbox. Для обработки данных используем СУБД Hive в среде Hadoop, позволяющую осуществлять операции над данными и их анализ путем SQL – подобных запросов. Для этого создаем файл обработки и создания нужных таблиц hivedll.sql. Вид файла показан ниже:

```
//описание идентификаторов таблиц из twitter
CREATE EXTERNAL TABLE tweets_raw (
id BIGINT,
created_at STRING,
source STRING,
favorited BOOLEAN,
retweet_count INT,
retweeted_status STRUCT<
text:STRING,
usr:STRUCT<screen_name:STRING,name:STRING>>,
entities STRUCT<
urls:ARRAY<STRUCT<expanded_url:STRING>>,
user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
hashtags:ARRAY<STRUCT<text:STRING>>>,
text STRING,
usr STRUCT< screen_name:STRING,  name:STRING, friends_count:INT, followers_count:INT, statuses_count:INT,
verified:BOOLEAN, utc_offset:STRING, -- was INT but nulls are strings time_zone:STRING>,
in_reply_to_screen_name STRING,
yearint,
monthint,
dayint,
hourint
)
CREATE EXTERNAL TABLE time_zone_map (
time_zone string,
country string,
notes string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
```

```
LOCATION '/user/data/time_zone_map';
…
create table tweets_sentiment stored as orc as select
id,
case
when sum( polarity ) > 0 then 'positive'
when sum( polarity ) < 0 then 'negative'
else 'neutral' end as sentiment
from l3 group by id;
 -- put everything back together and re-number sentiment
CREATE TABLE tweetsbi
STORED AS ORC
AS
SELECT
t.*,
cases.sentiment
when 'positive' then 2
when 'neutral' then 1
when 'negative' then 0
end as sentiment
FROM tweets_clean t LEFT OUTER JOIN tweets_sentiment s on t.id = s.id.
```

Запускаем данный файл командой: Hive_f hiveddl.sql. Структурированные данные будут размещены в таблице 1.

Таблица 1. Вид заголовков для анализа структурированных данных в задачах для МБ

A	B	C	D	E	F
Data/Time	Time/Zona	language	Text	location	Sentiments

Для анализа данных используются следующие показатели. Общее количество твиттов (Kol_i) для каждого места размещения (R) определяется:

$$Kol_R = \sum_{i=1}^{N} k_i, k_i \in R,$$

где k_i - каждый следующий твитт из обрабатываемого потока.

Частота употребления уникального слова ch(m) определяется из общего множества L текстовых данных:

$$ch(m) = \sum_{i=1}^{N} m_i, m_i \in L.$$

Отношение каждого твитта otn (m,rez) может быть определен из тезауруса tez, в котором прописано отношение к данному слову:

$$otn(m, rez) = \begin{cases} 0, m - негативное\_значение \\ 1, m - нейтральное\_значение \\ 2, m - положительное\_значение. \end{cases}$$

Для дальнейшей работы был составлен словарь, состоящий из фильтров предметной области, чтобы в дальнейшем определить количество твиттов по размещению ch(m) и количество твиттов по размещению с учетом отношения otn (m,rez). Определяем тезаурус с учетом фильтров по базовым факторам МСБ: пища, одежда, развлечения и дети. В итоге получаем 4 базовых фактора предпринимательства МБ.

По фактору «пища» P_1 получаем количество твиттов в общем множестве текстовых данных L:

$$Kol_{otnP_1} = \frac{\sum_{i=1}^{N} S_i (S_i \in P_1)}{L} = 9\%.$$

По фактору «одежда» P_2 получаем количество твиттов в общем множестве текстовых данных L:

$$Kol_{otnP_2} = \frac{\sum_{i=1}^{N} S_i (S_i \in P_2)}{L} = 8\%.$$

По фактору «развлечения» P_3 получаем количество твиттов в общем множестве текстовых данных L:

$$Kol_{otnP_3} = \frac{\sum_{i=1}^{N} S_i (S_i \in P_3)}{L} = 6\%.$$

По фактору «дети» P_4 получаем количество твиттов в общем множестве текстовых данных L:

$$Kol_{otnP_4} = \frac{\sum_{i=1}^{N} S_i (S_i \in P_4)}{L} = 12\%.$$

4. Результаты и обсуждение

В итоге можно сделать вывод о том, какая область МБ особенно востребована в Самарской области. По рисунку 3 видно, что основная стратегия продвижения МБ для органов власти должна быть связана с открытием детских центров.

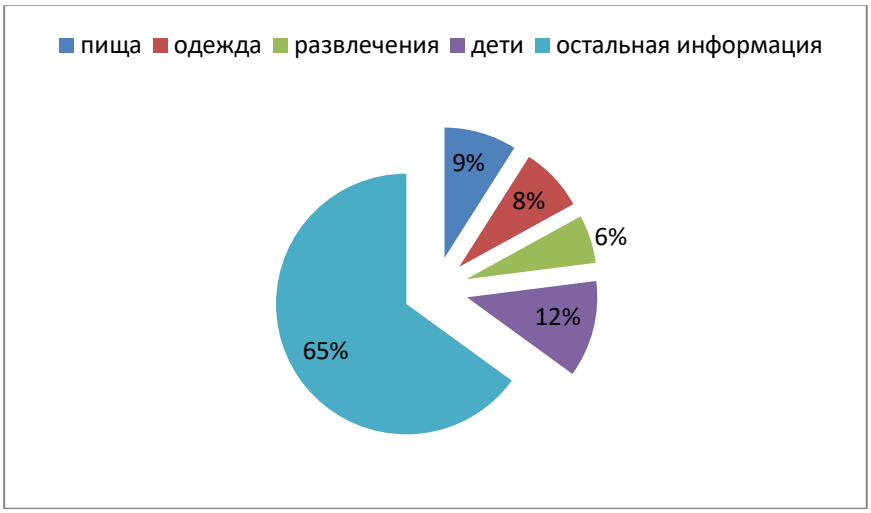

Рис. 3. Распределение факторов МБ в Самарской области.

Благодаря технологии BIG DATA можно хранить и обновлять данные в файловой системе «hadoop» по фильтру «Самарская область» (filter1= {Самарская область}). Затем необходимо данную область отфильтровать по базовым факторам малого и среднего предпринимательства, установив, например, следующие фильтры: Filter2 (пища) = {кафе, бар, ресторан, кух*, пивн*, мясо, рыба, трактир}; Filter3 (одежда)= {куртки, кофты, плать*, юбк*, кофт*, лиф*, шмотк*}; Filter4 (развлечения)= {ночной клуб, концерт, сейшн, тусовка}; Filter5 (дети) = {детсад, бэби-клуб, секция}.

Набор дескрипторов, по которым будет осуществляться фильтрация Интернет-дискурса определяется лексическими репрезентантами понятия, сформировавшегося в картине мира среднестатистического русскоязычного потребителя услуг. Основной в концептосфере «Еда» является микроситуация «Приготовление пищи», которая включает следующую когнитивно-пропозициональную структуру: Субъект - Предикат приготовления пищи (каким образом готовит) - Объект приготовления пищи - Свойство объекта приготовления пищи - Способ приготовления пищи - Помещение - Посуда - Приспособление - Прибор - Предприятие - Вещество - Пищевой продукт / кушанье - Свойство пищевого продукта / кушанья В ситуации Интернет-общения экспликацию получают лишь актуальные для пользователя элементы структуры, лексическое наполнение которых позволит нам сделать вывод о потребностях жителей конкретного района Самары. Формирование массива дескрипторов по Filter3 (одежда); Filter4 (развлечения); Filter5 (дети) может быть осуществлено с опорой на лексико-семантические поля «одежда», «мода»; ассоциативно-семантическое поле «отдых»; концепт «детство».

Для принятия решений в области МБ в регионе необходима мультимодальная кластеризация социальных сетей. Кластеризация основана на методе анализа формальных понятий (Formal Concept Analysis,FCA). Большое количество структурированных и неструктурированных данных генерирует тривиальные данные. Например, данные социальных сайтов в области МБ можно представить в виде следующей тройки (пользователь, группа, интерес) (рис. 4).

По методу формальных понятий вводим следующие определения: G - множество объектов, M – множество признаков, отношение $I \subseteq G \times M$ такое, что $(g,m) \in I$ тогда и только тогда, когда объект g обладает признаком m; $K := (G, M, I)$ называется формальным контекстом.

Определяем операторы Галуа следующим образом: для $A \subseteq G, B \subseteq M$ $A' \stackrel{def}{=} \{m \in M \,|\, g / m \forall g \in A\}$, $B' \stackrel{def}{=} \{g \in G \,|\, g / m \forall m \in B\}$, где A – формальный объем, B – формальное содержание.

Формальное понятие есть пара $(A,B): A \subseteq G, B \subseteq M, A' = B$ и $B' = A$.

Понятия, упорядоченные отношением $(A_1, B_1) \geq (A_2, B_2) \Leftrightarrow A_1 \supseteq A_2 (B_2 \supseteq B_1)$, формируют полную решетку, называемую контекстной решеткой $\underline{\beta}(G, M, I)$. Пример контекста социальных сетей в области МБ и их контекстная решетка показаны в таблице 2 и на рисунке 5.

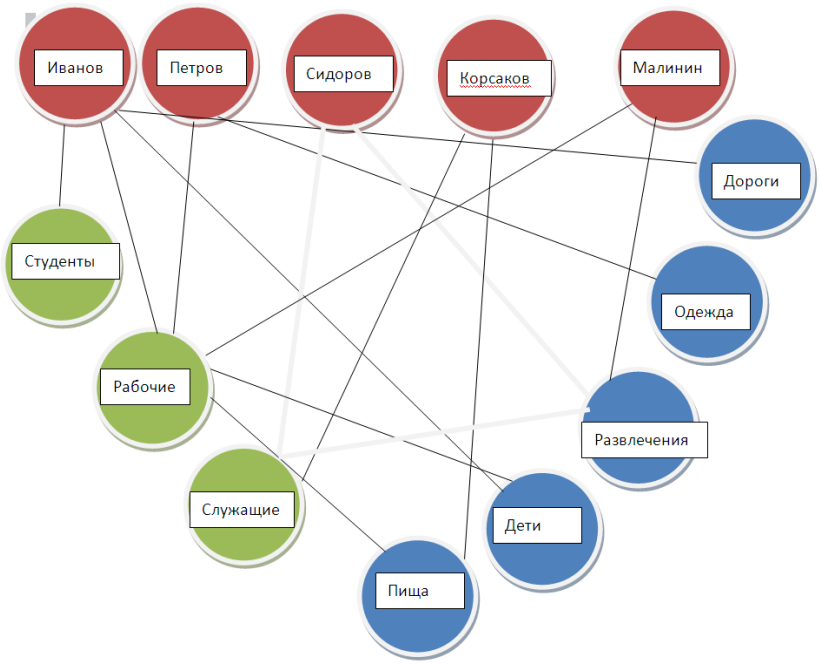

Рис. 4. Данные по МБ из социальной сети «twitter» как граф.

Таблица 2. Пример контекста данных по МБ из социальной сети (а - атрибуты по фильтру «пища», б - атрибуты по фильтру «дети», в - атрибуты по фильтру «развлечения», г -атрибуты по фильтру «одежда»)

	G/M	a	b	c	d
1	пенсионеры	x			x
2	служащие	x		x	
3	рабочие		x	x	
4	студенты		x	x	x

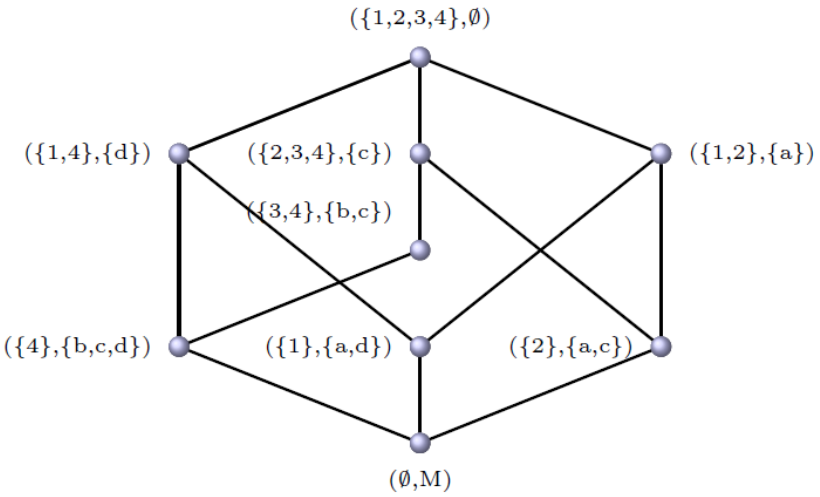

Рис. 5. Контекстная решетка для социальной сети.

Использование данного метода кластеризации позволит определить группы по интересам, с увеличением связей в которых потребуется принимать управленческие решения. Но данный инструмент имеет ограничения по использованию. Пользователи, которые работают с социальной сетью «Twitter», находятся в группе «студенты» и частично в группах «служащие», «рабочие» и лишь незначительно затрагивают часть группы «пенсионеры», поэтому для полноты принятия управленческих решений необходимо добавлять полевые маркетинговые исследования в указанных группах.

Можно получить графики зависимости количества обращений пользователей по фильтрам от времени сбора данных [9]. Время сбора данных из Интернета в технологии BIG DATA неограниченно.

В итоге получаем динамическое изменение информации в режиме реального времени из системы Интернет, что позволяет с минимальными инвестициями проводить мониторинг потокового анализа неструктурированной информации

(технология In-Memory Data Processing and Stream) по фильтрам. Для реализации данного метода была написана программа на языке Scala:

```
val file = spark.textFile(“hdfs://… “)
val errors=file.filter(line=>line.contains(“Самарская область“))
//count all the data
errors.count()
//count data mentioning Filter
errors.filter(line=>line. contains(“мясо“)).count()
//Fetch the filter as an array of string
errors.filter(line=>line. contains(“пища“)).collect()
```

После работы программы получаем динамическое изменение параметров в среде BIG DATA, которые позволяют определять зоны малого бизнеса в регионе с учетом неструктурированной информации. В случае выявления на графиках устойчивых «всплесков» данных по количеству обращений в соответствии с формами предпринимательства должна осуществляться инвестиционная поддержка по развитию малого и среднего бизнеса по данному виду деятельности в рассматриваемой зоне.

Таким образом, предложен инструмент для повышения эффективности использования бюджетных средств в регионе. Это является важнейшей задачей в современных экономических условиях, в основе решения которой лежит возможность принятия оптимальных управленческих решений. Предложенный способ регулирования может быть эффективен при управлении различными инновационными процессами развития экономики региона, для которых характерны многообразие форм и широкий спектр составляющих и факторов, а также свойственна динамика развития и активная трансформации жизнедеятельности.

При этом использование современных программно-аппаратных средств позволяет производить оценку и визуализацию изменений фактически в режиме реального времени, что может быть полезно не только органам власти на местах, но и бизнесу в процессе разработки и реализации инвестиционных проектов.

Литература

[1] Дровянников, В.И. Разработка модельного аппарата управления конкурентным развитием социального кластера региона/ В.И. Дровянников, И.Н. Хаймович//Фундаментальные исследования. -2015.- №7(ч. 4).-С. 822-827.

[2] Дровянников, В.И. Имитационное моделирование управления социальным кластером в системе Any Logic / В.И. Дровянников, И.Н. Хаймович//Фундаментальные исследования. -2015.- №8 (ч.2). - С. 361-366.

[3] Рамзаев, В.М.Разработка модели функционирования производственных активных элементов в региональном управлении/ В.М. Рамзаев, Е.А. Кукольникова, И.Н. Хаймович// Вестник СГЭУ. -2014. - №12. – С.87-99.

[4] Рамзаев, В.М. Комплексная модель управления экономическим развитием региона на основе повышения конкурентоспособности предприятий/ В.М. Рамзаев, И.Н. Хаймович// Современные проблемы науки и образования. – 2014. – № 6.-С.136.

[5] Рамзаев, В.М. Модели прогнозирования конкурентного роста предприятий при энергомодернизации / В.М. Рамзаев, И.Н. Хаймович, В.Г. Чумак// Проблемы прогнозирования. 2015.- №1. – С. 67-75.

[6] Bonacich, P. Power and Centrality: A Family of Measures/ P. Bonacich//American Journal of Sociology.-2007. – V. 92(5). – P.1170-1182.

[7] Chumak, P.V. Models for forecasting the competitive growth of enterprises due to energy modernization/ P.V. Chumak, V.M. Ramzaev, I.N. Khaimovich//Studies on Russian Economic Development. – 2015.- V.26.No.1.- P. 49-54.

[8] Chumak, V.G. Challenges of Data Access in Economic Research based on Big Data Technology/ P.V. Chumak, V.M. Ramzaev, I.N. Khaimovich// CEUR Workshop Proceedings. -2015. – V.1490.- P. 327-337.

[9] Chumak, V.G.Use of Big Data technology in public and municipal management/ P.V. Chumak, V.M. Ramzaev, I.N. Khaimovich//CEUR Workshop Proceedings. -2016. – V.1638..- P.864-872.

[10] Khaimovich , A.I. Development of the requirements template for the information support system in the context of developing new materials involving Big Data/ F.V. Grechnikov, A.I. Khaimovich//CEUR Workshop Proceedings.-2015. – V.1490. -P.364-375.