Знания-Онтологии-Теории (ЗОНТ-09)

Исследование американских политических блогов на основе формального анализа понятий

Михаил Климушкин, Дмитрий Четвериков
Государственный Университет – Высшая Школа Экономики, ул. Кирпичная, д. 33/5, г. Москва, Россия
klim.mikhail@gmail.com, dmchetverikov@gmail.com

Аннотация. *В данной статье описываются исследования данных об американских политических блогах. Целью нашего исследования было построение структуры интересов блогеров, выявление основных тем и отслеживание изменений в этой структуре со временем. Сложность исследования заключается в большом количестве шума, так как блоги пишутся в свободной форме, и блогеры используют слова, не относящиеся к основной теме блога.*
Основным инструментом исследования является формальный анализ понятий (ФАП). Ниже описываются основные термины и методы, которые использовались для анализа. ФАП позволил построить основную структуру обсуждаемых тем, отсевая шум, а также увидеть изменение интересов групп блогеров со временем.

Ключевые слова: формальный анализ понятий, обнаружение закономерностей, бикластеризация

1 Введение

Цель нашей работы заключалась в применении методов формального анализа понятий к обработке информации, полученной по политическим блогам США. Средства ФАП использовались ранее в анализе информации о посещаемости интернет ресурсов, структуры аудиторий сайтов и для выделения различных групп среди целевой аудитории. Эта статья является попыткой применить такие методы над политическими данными. Преимуществом данных методов является наглядное и удобное для изучения представление результатов в виде решеток. Так как данные предоставлены за период, в котором шла предвыборная гонка, одной из наших задач стало определение наиболее обсуждаемых политиков и изменение состава лидеров этой гонки. В отличие от статистических методов, формальный анализ понятий позволяет строить структуры интересов блогеров в виде решеток формальных понятий, что дает возможность наглядно показать всю структуру интересов[1]

2 Данные

Данные были предоставлены кампанией RTGI. Эта французская аналитическая кампания занималась анализом американских политических блогов. Для анализа из текстов блогов были выбраны 79 слов, которые, по мнению специалистов RTGI, должны отражать тематики блогов. Затем были собраны данные за период с 1 ноября 2007 по 29 мая 2008 года.

Так как данные имеют временную характеристику, мы можем не только построить тематическую структуру политических блогов, но и проследить изменения этой структуры со временем.

Основным методом анализа блогов будет формальный анализ понятий, создание формальных контекстов и построение по ним решеток формальных понятий.

[1] Данное исследование было поддержано Научным Фондом Государственного Университета – Высшей Школы Экономики (№ 08-04-0022).

3 Создание формального контекста

Формальный контекст – это тройка *K=(G,M,I)*
G – множество объектов
M – множество признаков
I – отношение обладания признаком. $I \subseteq G \times M$
$I = \{\langle g_i, m_j \rangle\}$. Пара $\langle g_i, m_j \rangle$ показывает, что объект g_i обладает признаком m_j
Часто формальный контекст представляют в виде бинарной матрицы:

Таблица 1. Пример формального контекста

	m1	m2	m3	m4
g1			X	X
g2		X	X	
g3	X			X
g4	X	X	X	

В нашем случае объектами будут блогеры, а признаками – 79 ключевых слов.

Прежде всего, выберем временной период (день, неделю). Для каждого блогера посчитаем, сколько раз за этот период блогер употреблял каждое слово, т.е. мы получаем матрицу объекты-признаки *Q*, в которой элемент q_{ij} равен количеству употреблений i-м блогером j-го слова.

Из данной матрицы необходимо получить бинарную. Для этого мы устанавливали порог на количество употреблений слова. Если блогер использовал слово больше заданного порога, считаем, что блогер активно использовал это слово, следовательно, обсуждал тему, к которой относится это слово. Порог необходим, чтобы отсечь случаи случайного употребления слова.

4 Построение решетки формальных понятий

Теперь имея контекст, необходимо построить структуру интересов, выделить группы блогеров со схожими интересами. Для этого построим решетку формальных понятий.

В определении формального понятия используется операторы Галуа:

Для $A \subseteq G$ и $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A: gIm\}$$

$$B' = \{g \in G \mid \forall m \in B: gIm\}$$

Другими словами *A'* – множество признаков, которыми обладают все объекты из множества *A*. *B'* – множество объектов, которые обладают всеми признаками из множества *B*.

Формалинное понятие (A,B) состоит из множества объектов $A \subseteq G$ и множества признаков $B \subseteq M$, таких что $B'=A$ и $A'=B$. *A* называется объемом, а *B* – содержанием понятия. В матрице контекста формальное понятие представляет собой подматрицу, состоящую и единиц:

Пример, формальное понятие *([g2,g4],[m2,m3])*:

Таблица 2. Пример формального понятия

	m1	m2	m3	m4
g1			X	X
g2		X	X	
g3	X			X
g4	X	X	X	

Множество понятий контекста *K* образуют решетку формальных понятий $\beta(K)$, так как создают частичный порядок по вложению объемов понятий и всегда имеют наименьшее и наибольшее по вложению понятия.

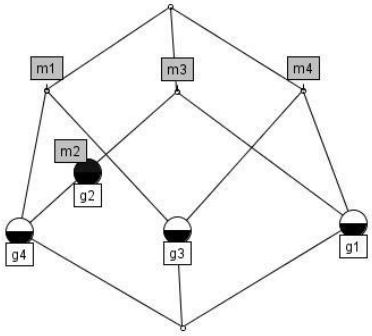

Рис 1. Пример решетки формальных понятий

Как читается эта решетка? Каждая вершина решетки – формальное понятие. Рядом с понятием пишутся объекты, которых нет в менее общих понятиях (находящихся под данным понятием), и признаки, которых нет в более общих понятиях. Тогда объем формального понятия – все объекты, написанные напротив данного понятия и всех понятий, менее общих, чем оно. Содержание – признаки, написанные напротив данного понятия и более общих понятий.

Например, понятие с подписями «m2» и «g2» имеет объем [g2,g4] и содержание [m2,m3] .

Находятся такие формальные понятия алгоритмом «замыкай по одному». Функция начинает работать с самого общего формального понятия, которое содержит все объекты и чаще всего ни одного признака. Затем находятся все остальные понятия рекурсивным добавлением признаков.

Но использовать всю решетку формальных понятий не всегда удобно из-за ее громоздкости. Количество формальных понятий экспоненциально зависит от размера матрицы. Например, контекст, составленный по данным за 1 день (1 ноября), состоит из 49 блогеров и 65 слов. Решетка, построенная по данному контексту, имеет 202 формальных понятия:

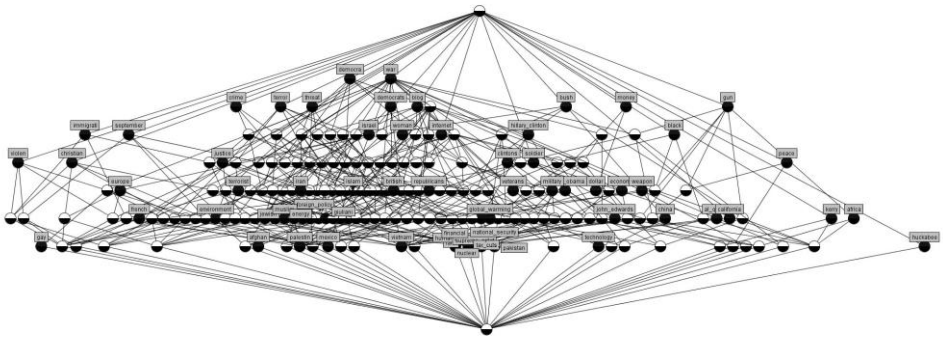

Рис 2. Решетка формальных понятий за 1 ноября 2007

Контексты за более большие периоды (например, неделю) могут содержать более 1 млн. понятий. Конечно, анализировать их не представляется возможным.

Решетка громоздкая, но многие формальные понятия не несут практически никакой информации или возникли из-за шума. Поэтому рационально пренебречь малозначимыми формальными понятиями для получения более простой решетки.

Один из способов отбора наиболее важных понятий - индекс устойчивости.

5 Индекс устойчивости

Различают два типа устойчивости: интенсиональную и экстенсиональную.

Экстенсиональная устойчивость показывает, насколько объем формального понятия зависит от отдельного признака. Индекс высчитывается по формуле:

$$\sigma(A,B) = \frac{\left|\{D \subseteq B : D' = A\}\right|}{2^{|B|}}$$

Интенсиональная устойчивость, показывает, насколько содержание формального понятия зависит от отдельного объекта.

$$\sigma_{in}(A,B) = \frac{\left|\{C \subseteq A : C' = B\}\right|}{2^{|A|}}$$

Индекс равен доли подмножеств множества объектов, порождающих данное формально понятие.

Так как нас в первую очередь интересует содержания понятий, то есть ключевые слова, то мы будем использовать интенсиональную устойчивость и отбирать понятия с наиболее устойчивыми множествами слов.

6 Исследование блогов

Теперь мы можем находить все формальные понятия, выбирать наиболее устойчивые из них, и получить более компактные решетки, которые при этом сохраняют основную структуру интересов.

Например, выберем из решетки понятий, построенной по данным за 1 ноября, понятия, с интенсиональной устойчивостью более 0.9. Получаем вполне читабельную и удобную для анализа решетку[2]:

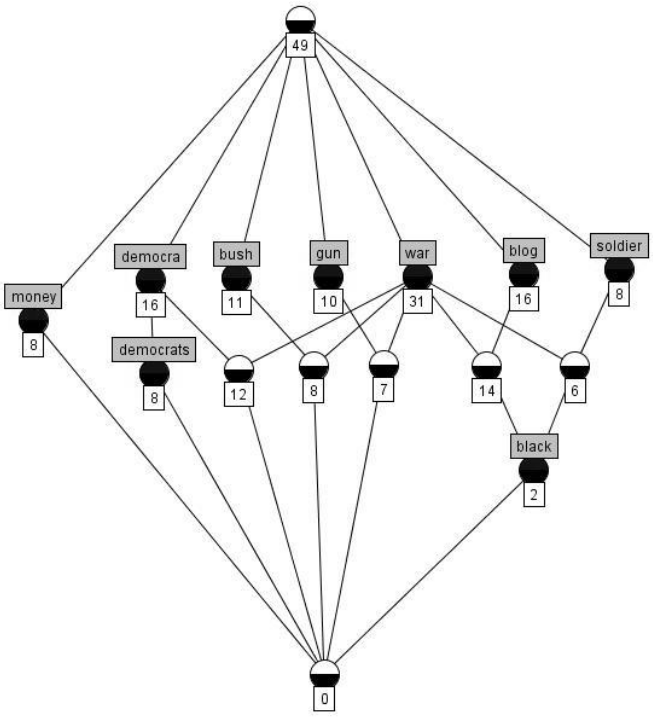

Рис 3. Решетка с устойчивыми понятиями за 1 ноября 2007

Можем заметить, что самое популярное слово – ‘war’. Видимо в этот день активно обсуждались военные действия, которые велись в то время в мире. Так же с войной связано

[2] В этой и последующих решетках числа под понятиями означают количество блогеров в данном понятии

слово ‘bush’, видимо эти люди обсуждали войны с участием США и внешнюю политику, которую вел Буш.

Теперь построим решетку за сотый день, когда блогеры были более активны:

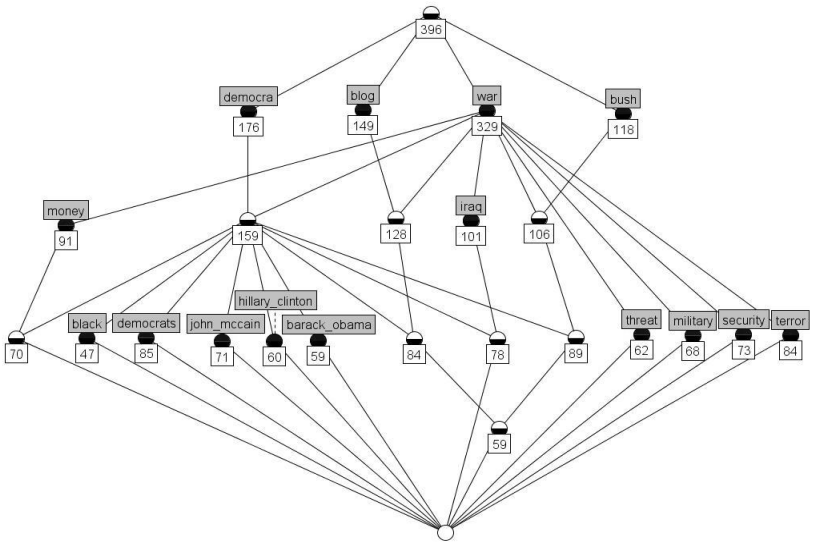

Рис 4. Решетка формальных понятий за 8 февраля 2008

По решетки видно, что главным образом в блогах наиболее популярны 2 области: война (правая часть решетки) и тема предстоящих выборов (левая часть решетки)

Для того, что бы посмотреть как изменялись предпочтения блогеров во время предвыборной гонки, оставим в нашем контексте только имена кандидатов на пост президента США. Таких среди 79 слов семь: Барак Обама, Хилари Клинтон, Джон Эдвардс, Джон Маккейн, Митт Ромни, Руди Джулиани и Майк Хакаби.

Разобьем наши данные на 17 периодов по 7 дней. Составим контексты по каждому периоду и оставим в них только по 7 признаков, имен политиков. Построим решетки по каждому из 17 периодов. По ноябрьским решеткам видно четкое разделение политиков на демократов и республиканцев:

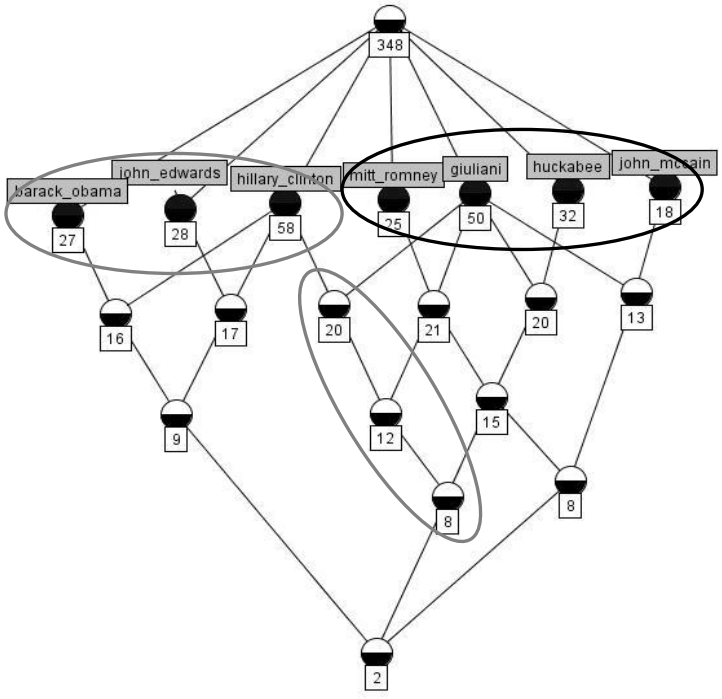

Рис 5. Обсуждения политиков, ноябрь

При этом можно заметить, что в отличие от других демократов Хилари Клинтон имеет много общих с республиканцами формальных понятий. Отсюда можно предположить, что в начале предвыборной гонки Клинтон была наиболее ожидаемым претендентом от демократов. При этом среди республиканцев наиболее популярен Джулиани.

Наиболее сильные изменения в структуре произошли на первой неделе февраля, резко упал интерес к Джулиани и Эдвардсу, и в середине февраля, упал интерес к Митт Ромни:

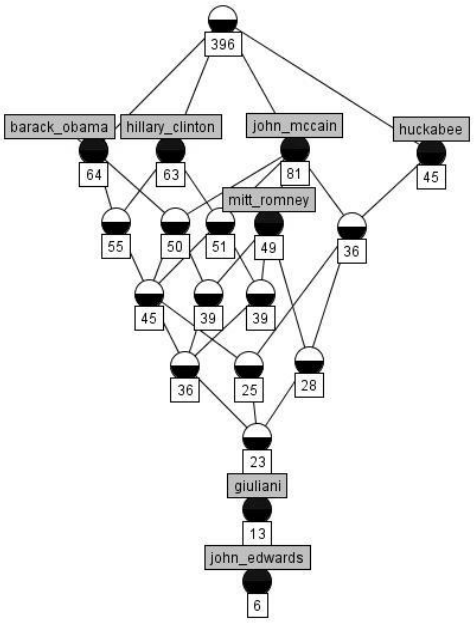

Рис 6. Обсуждения политиков, начало февраля

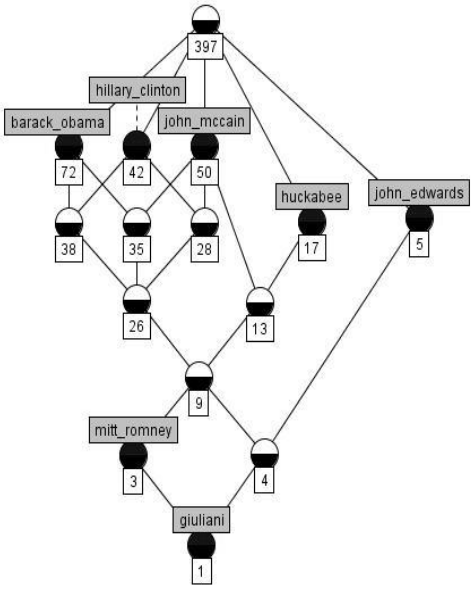

Рис 7. Обсуждения политиков, середина февраля

Причиной такого изменения стал выход этих политиков из предвыборной гонки.

При этом явно видно, что среди оставшихся претендентов Обама, Клинтон и Маккейн довольно активно обсуждаются блогерами, в то время, как Хаккаби гораздо менее популярен, т.е. можно предположить, что Хаккаби следующим закончит гонку. Как показывает история, он закончил свою предвыборную кампанию 4 марта 2008, через месяц.

Таким образом, по решеткам возможно с точностью до недели сказать, когда какой политик выбыл из гонки, но возникает вопрос, какого политика стали поддерживать избиратели, поддерживавшие выбывших политиков?

Для того чтобы это понять, сравним контексты за период, в котором политик еще участвовал в предвыборной гонке и за период, когда он выбыл из нее.

Для сравнения будем использовать экстенсиональную связность.

7 Экстенсиональная связность

Рассмотрим формальное понятие *(A,B)* из одного контекста и формальное понятие *(C,D)* из второго контекста.

Возьмем пересечение объемов понятий. Если замыкание этого множества в первом контексте равно *A*, а во втором – *C*, то эти формальные понятия экстенсионально связаны.

$$\begin{cases} K_1 : (A \cap C)'' = A \\ K_2 : (A \cap C)'' = C \end{cases}$$

Введем также порог на размер множества общих объектов. Считаем, что формальные понятия экстенсионально связаны если размер пересечения их объемов превышает заданной доли от размера объемов (например, 0.5 от объема понятия) Таким образом, сравниваются все понятия первого контекста со всеми понятиями второго.

Сравним контексты за период, в котором политик еще участвовал в предвыборной гонке и за период, когда он выбыл из нее.

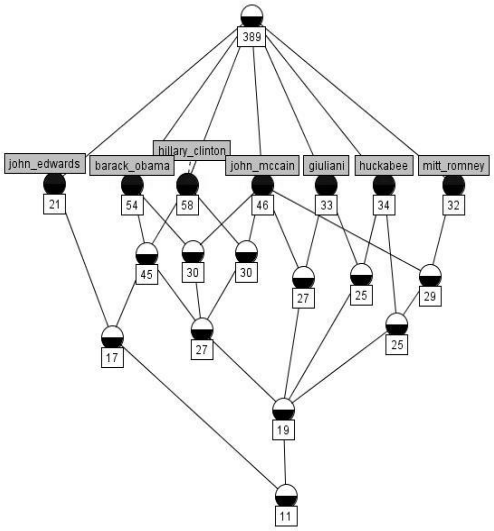

Рис 8. Обсуждения политиков, январь

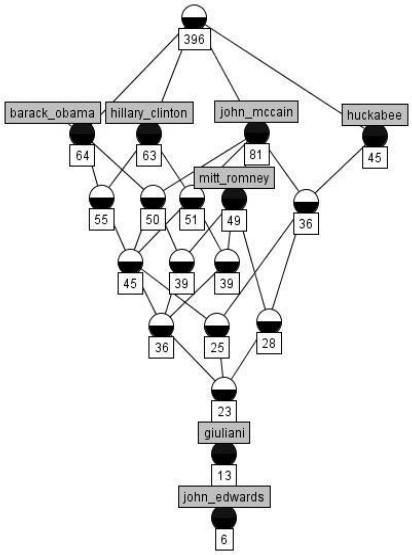

Рис 9. Обсуждения политиков, середина февраля

Получаем, что формальное понятие с содержанием ‘giuliani’ экстенсионально связано с понятиями со словом ‘john_mccain’. А понятие со словом ‘john_edwards’ – с понятиями со словом ‘barack_obama’. Следовательно, можно сделать вывод, что избиратели, поддерживавшие Джулиани, стали поддерживать Маккейна, а поддерживавшие Эдвардса, стали поддерживать Обаму. Как показывает история, Джулиани советовал своим избирателям поддерживать Маккейна, что подтверждает результаты, полученные формальным анализом понятий.

8 Оптимальные наборы признаков

Кроме индекса устойчивости можно попробовать сократить сам контекст. Для этого мы можем использовать уже готовые методы оптимизации, возникшие в разных научных областях. Но, во-первых, необходимо определиться по каким критериям следует отбирать признаки (далее будет подразумеваться именно выбор оптимального набора признаков, в нашем случае это слова, которые употребили блогеры). Например, можно искать такой минимальный набор признаков, которыми будут обладать все объекты, поскольку слова такого набора, как можно ожидать, будут играть ключевую роль в описании всего сообщества. Так же необходимо, чтобы пересечение по объектам у них было наименьшее.

Основываясь на этих критериях важности, были реализованы и апробированы оптимизационные методы такие как: метод Монте-Карло, метод наибыстрейшего спуска, генетический алгоритм. В качестве эксперимента, было реализован выбор 20-ти признаков из решетки ФП по американским блогам за весь период. Результатом такой выборки стали следующие слова: 'democra','wall_street','dollar','islam','financial','iran','afghan','democrats','september','national_security', 'money','george_w_bush','violen','mexico','pakistan','huckabee','immigrati','mitt_romney','energy','palestin'.

Если теперь рассмотреть решетку, состоящую только из этих 20 признаков, а объекты оставить без изменения, то уже можно значительно быстрее выделить получить все формальные понятия из таких данных и построить диаграмму решетки ФП.

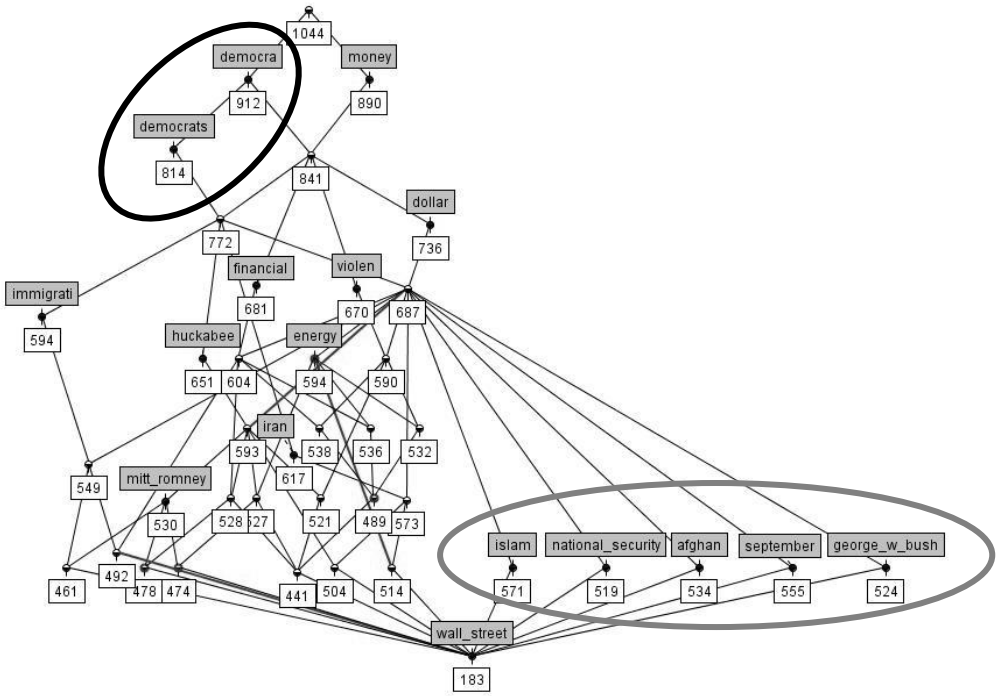

Рис 10. Решетка с выбранными признаками

По такой диаграмме уже сокращенной решетки формальных понятий, мы можем, например, проверить адекватность упомянутого подхода к редуцированию данных. Например, как видно из диаграммы, большинство блоггеров, говоря о демократах, упоминали также слово демократия. Аналогично можно сказать и про доллар, о котором говорили в контексте с деньгами. При детальном изучении диаграммы мы видим, что слова «ислам», «национальная безопасность», «афган», «сентябрь», «Джордж Буш» употребили из 687 человек приблизительно по 500 человек, и как минимум 183 человека высказали их в одном контексте. Иными словами, теперь, отталкиваюсь от полученных результатов, можно попробовать сконцентрироваться на проверке гипотезы: «правда ли, что Дж. Буша обсуждали только в контексте его политики в области национальной безопасности и результатов его работы в этой сфере». Слово 'huckabee' зачастую употреблялось вместе с 'money' и 'dollar'. Можно предположить, что блогеры интересовались предложенными кандидатом вариантами вывода США из кризиса, связанного с вливанием денег в экономику государства.

9 Метод k-средних

Для работы с политическими блогами, так же можно применить метод k-средних к проблеме оптимизации данного набора признаков. Все признаки разбиваются на кластеры. Затем, создается контекст, в котором в качестве признаков используются полученные кластеры. Если блогер использовал больше половины слов из кластера, то он обладает признаком, соответствующем данному кластеру.

В качестве эксперимента, мы выделяли за различные периоды времени по 20 кластеров из набора признаков. В качестве названия кластера использовалось одно из слов кластера. Были получены кластеры:

[cluster's name]: [words in the cluster]

pakistan: afghan pakistan
september: nuclear september
palestin: israel palestin
technology: british crime environment internet technology
super_tuesday: democrats primaries republicans super_tuesday

french: europe french middle_east military security soldier terror terrorist threat violen weapon

al_qaeda: al_qaeda dollar iran islam muslim
supreme_court: gun justice supreme_court
и др.

Теперь построим решетку формальных понятий, отбирая наиболее устойчивые[3]:

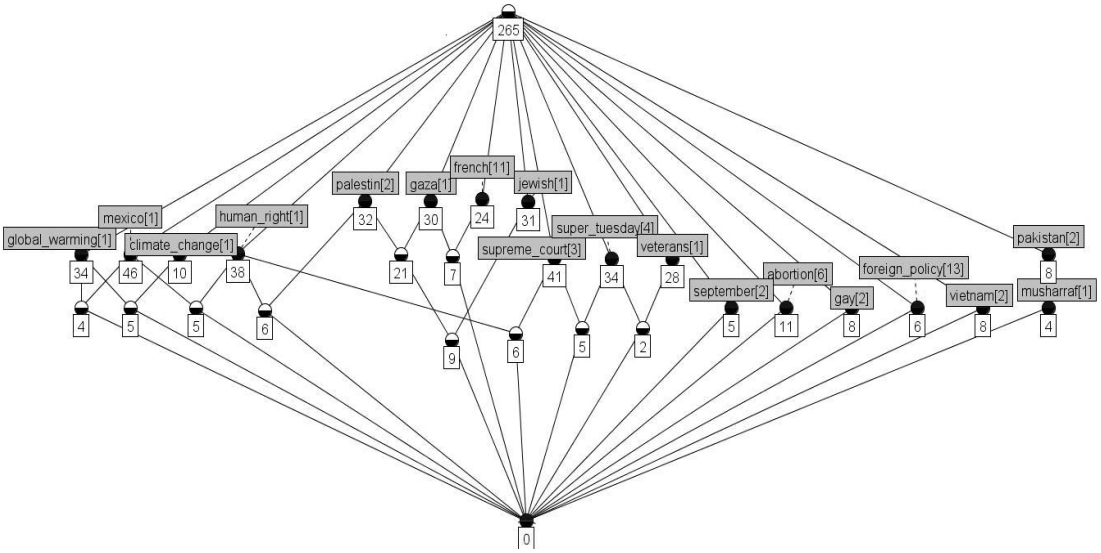

Рис 11. Решетка с группами признаков

После таких преобразований, данные по политическим блогам легче поддаются анализу. Из диаграммы мы видим формальные понятия с содержанием [«Mexico», «human_right»] и [«Palestin», «human_right»]. Можно рассмотреть такую гипотезу: «участников американских блогов если и интересуют права человека, то в Мексике и Палестине». Также мы видим связь кластера “french” ,“gaza” и “palestin”, которая, как можно предположить, отвечает за обсуждения ближневосточных проблем и участия Европы в их решении.

10 Выводы

В этой работе мы попытались продемонстрировать возможность применения методов формального анализа понятий к данным, полученным по политическим блогам. В качестве результата мы получили демонстрацию наглядного представления процессов протекающих в политике в ходе выборов на пост президента в США. Наглядное представление обширных данных, полученных по блогам, может помочь эксперту концентрировать внимание на наиболее интересных фактах и событиях. Основываясь на полученных результатах, были также предприняты попытки спрогнозировать дальнейшие пути развития событий, которые подтвердились в дальнейшем. Кроме этого, был представлен метод, позволяющий выявлять связи между политиками, направлениями их работы и проводить оценку таких связей.

Литература

[1] Vilem Vychodil: A New Algorithm for Computing Formal Concepts. Binghamton, University - SUNY, Binghamton, USA, 2008.

[2] Camille Roth, Sergei Obiedkov, Derrick Kourie: Towards Concise Representation for Taxonomies of Epistemic Communities

[3] Kuznetsov, S.O.: On stability of a formal concept. France, 2003

[3] В квадратных скобках после названия кластера пишется количество слов в данном кластере