НЕКОТОРЫЕ МЕТОДЫ АНАЛИЗА НОВОСТНЫХ ИНФОРМАЦИОННЫХ ПОТОКОВ

Ландэ Д.В.
Информационный центр "Элвисти", Киев

Abstract

Lande D. Some methods of news streams analysis. *Different methods of clusterization for news streams analysis are described. Best method for system InfoStream® is proposed.*

Введение

Сегодня Web-пространство охватывает свыше 20 млрд. документов, и темпы роста этого информационного ресурса превращают его в информационный поток [1]. Исследование новостной составляющей этого потока, т.е. потока новостных сообщений, публикуемых на страницах Web-сайтов, должно использовать принципиально новый инструментарий, так как классические методы кластерного анализа не всегда способны адекватно отражать ситуацию. В этом случае речь идет не столько об анализе документального массива фиксированной размера, сколько о навигации в потоке документов. В настоящей публикации приведены некоторые методы, которые, по мнению автора, могут эффективно применяться для анализа информационных потоков.

1. Подход TF*IDF для информационного потока

Любая модель, в том числе, модель информационного потока предполагает некоторое упрощение, предположим, учитывая это обстоятельство, что информационный поток – это динамично меняющееся множество документов, которые, в свою очередь, состоят из термов – слов и устойчивых словосочетаний. Введем обозначение информационного потока:

$$D(t) = \{Di, i=1,..,N(t)\}$$

где Di – документ с номером i, t –время, N(t) – количество актуальных документов в потоке в момент t. Di={wij}, где wij - множество термов, входящих в документ Di. Предполагается, что новостные документы устаревают, теряя свою актуальность. Если предположить, что ранг актуальности в момент публикации документа равен 1 и λ - коэффициент полураспада ранга актуальности, т.е., что $e^{-\lambda \Delta t}$

= ½, где Δt – промежуток времени (в часах), за который документ в информационном потоке, который ввиду устаревания теряет свою актуальность наполовину. Например, если предположить, что документ в некотором тематическом новостном потоке за сутки теряет половину своей актуальности, то имеем: e-λ*24 = ½, и, соответственно, λ ≈ 0,029. Актуальным может считаться документ, у которого ранг актуальности превышает заданное заранее пороговое значение, например, 0.001.

Далее рассмотрим понятие информационного портрета. Портрет в широком понимании можно рассматривать как модель реального объекта, выраженную его наиболее узнаваемыми чертами. Используем понятие информационного портрета [2] как набора термов, а также некоторых весовых коэффициентов этих термов. Пусть тематической рубрике соответствует ее информационный портрет:

Pi = { pij , vij} (j=1,..,Ki)

где pij - терм, vij – соответствующий весовой коэффициент, Ki - количество термов (реальной тематической рубрике может соответствовать несколько сотен термов).

В данной статье предполагается, что информационные портреты зафиксированы, хотя в действительности, как состав информационных потоков, так и состав входящих в них термов, во многом зависят от самих потоков и не являются статичными объектами.

Пусть Ml = {mlij} (i = 1,..,N(t); j = 1,..,K) - матрица соответствия потока документов D(t) информационному портрету l. Определим в качестве значений элементов mlij весовые коэффициенты TF * IDF из классической пространственно-векторной модели [3]. Напомним, TF – это локальная частота терма (Term Frequency), а IDF – величина, обратная частоте встречаемости во всем потоке документов, содержащих данный терм (Inverse Document Frequency).

В то время как локальная частота терма в документе говорит о значимости терма в пределах документа, то обратная частота встречаемости свидетельствует об уникальности терма во всем потоке документов. Поэтому произведение этих величин – достаточно удачный критерий определения веса терма.

Примем такое расширенное толкование значения локальной частоты терма из информационного портрета l:

TF = log(1+ Vl * Npl/|Di|))

где Npl – количество экземпляров данного терма в документе Di, а Vl – вес этого терма из информационного портрета l. Таким образом, локальная частота – это вес терма из информационного портрета l в документе i. Очевидно, что если в документе один терм встретился три раза, а другой терм - один раз, то это по смыслу не следует, что первый терм в три раза весомее. Для сглаживания этого эффекта принято использовать логарифмирование, кроме того, для обеспечения

нормирования по удельному весу термов в документах разного размера используется размер документа – количество входящих в него термов |Di|.

Обратная частота встречаемости IDF обычно вычисляется по формуле:

IDF=log(1+N(t) / Ndlj)

Где N(t) – количество документов в D(t),

Ndlj - количество документов в D(t), которые включают данный терм из информационного портрета l.

Таким образом:

mlij = log(1+ Vl * Npl/|Di|)) * log(1+N(t) / Ndlj)

Естественно, в этой формуле элемент mlij равен нулю, если в документе нет соответствующего терма из информационного портрета.

Вес документа Di в пространстве тематической рубрики l определяется как нормированная сумма строки матрицы Ml , соответствующей этому документу:

VDil = 1/Kl * Σi log(1+ Vl * Npl/|Di|)) * log(1+N(t) / Ndlj)

При определенных заранее информационных портретах тематических рубрик, для любого документа могут быть вычислены его веса в пространстве всех тематических рубрик.

Также может быть введено минимальное значение веса документа по отношению к тематической рубрике, достаточное для рубрикации и отображения. Тематическая рубрикация может проводиться как для отдельных документов, так и для массива - результата поиска по запросу, но в этом случае имеет смысл наряду с весом каждой рубрики отобразить и количество релевантных документов.

Особый интерес представляют документы с низкими весами по всем информационным портретам, такие документы можно отнести к аномалиям и они зачастую могут порождать новые тематические рубрики.

Выявление принадлежности отдельного документа к различным рубрикам особенно эффективно и полезно при анализе их тематической направленности в случае, если они обладают большими размерами.

Процедура взвешивания потока документов в пространстве информационных портретов может выполняться в соответствии со следующим алгоритмом:

```
while не исчерпался список актуальных на момент t документов do
for каждого документа do
        while не исчерпался список информационных портретов
            for каждого информационного портрета do
                Определение веса документа
                if вес больше порогового значения
                then do приписывание документу рубрики,
                соответствующей информационному портрету
                end if
```

```
                end for
            end while
        end for
end while
    Итоговый подсчет веса потока документов
    Визуализация гистограммы
```

2. Взаимосвязь термов в информационном портрете

Итак, операция отображения потока документов в пространство информационного портрета l, задается матрицей M^l. Введем понятие ядра этой операции как произведения матриц $A = M^{lT} * M^l$[4]. Матрица A по смыслу представляет собой матрицу взаимосвязей термов в информационном портрете. При этом взаимосвязи рассчитываются на основании учета вхождения термов из информационных портретов в документы.

Еще одна матрица, полученная в результате умножения $B = M^l M^{lT}$ выражает взаимосвязь документов через присутствующие в них термы. Для современных информационных потоков размерность матриц этого типа намного превышает размерность матриц A. Вместе с тем, матрицы A и B связаны общим «происхождением», поэтому, например, выявляя явные группы взаимосвязанных термов в матрице A, можно предположить, что им будут соответствовать группы взаимосвязанных документов в матрице B, кластеризация которой в виду ее размерности и динамике роста весьма проблематична.

Наиболее взаимосвязанные термы, которые могут группироваться, можно выделять, например, перенумеровывая их, одновременно переставляя в матрице A соответствующие строки и столбцы. Таким образом, задача группировки взаимосвязанных термов сводится к задаче выделения диагональных блоков в матрице путем операции одновременной перенумерации соответствующих строк и столбцов. Такая блочная группировка матрицы позволяет наглядно определять новые семантические связи (возможно новые рубрики) на основании исходного информационного портрета. С другой стороны, учитывая корреляцию матриц A и B, эти новые семантические связи позволяют выделять массивы связанных документов, например, для формирования фрагментов распределенных баз данных.

Вместе с тем, поскольку речь идет об информационном потоке, а не о статическом массиве документов, перемножать матрицы M^{lT} и M^l на практике не так уж и просто, поэтому для анализа матрицы A предпочтительно ее динамическое формирование, а не явное перемножение матриц. Ниже приведен один из алгоритмов такого заполнения матрицы A

```
Определяется и инициализируется значениями 0 массив A размерностью
K*K
while не исчерпался список документов, зафиксированный в момент t  do
      for текущего документа do
            while не исчерпался список термов из информационного
портрета P_l do
                  for каждого терма p_i из P_l  do
                        if   вхождение терма p_i в документ
                        then do
                              Приписывание документу терма i
                              while список термов приписанных документу не
                              исчерпался  do
                                 if документу  приписан терм  j
                                 then увеличение элементов массива A[i,j] и A[j,i] на
                              m^l_ij *m^l_ji
                              end while
                        end if
                  end for
            end while
      end for
end while
Визуализация таблицы
```

4. Латентное семантическое индексирование

Покажем, как с целью выявления группы наиболее взаимосвязанных термов применяется метод кластерного анализа LSI (латентного семантического индексирования), который базируется на сингулярном разложении матриц [5]. Сингулярным разложением матрицы A (Singular Value Decomposition, SVD-разложением) называется ее разложение вида $A=USV^T$, где U и V – ортогональные матрицы, а S – диагональная матрица, элементы которой $s_{ij} = 0$, если i не равно j, а $s_{ii} >= 0$. Величины s_{ii} называются сингулярными числами матрицы A.

В рассматриваемом примере матрица $A = M^{lT} * M^l$ – квадратная, однако метод LSI применяется и к прямоугольным матрицам, но в этих случаях размерность матрицы S соответствует рангу матрицы A.

В соответствии с методом LSI в рассмотрение берутся k наибольших сингулярных значений, а каждому такому сингулярному значению матрицы A соответствует кластер взаимосвязанных термов информационного портрета. Таким образом матрица A аппроксимируется матрицей $A_k = \Sigma u_i s_{ii} v_i^T$, где $0<i<k+1$.

Известно, что A_k является лучшей аппроксимацией ранга k матрицы A в соответствии с нормой Фрбениуса, равной, сумме квадратов сингулярных значений.

В результате, в рассматриваемом случае, метод латентного семантического индексирования позволяет выделить из сотен термов информационного портрета несколько кластеров.

Метод LSI применим и к ранжированию выдачи информационно-поисковых систем, основанному на цитировании. Это алгоритм HITS (Hyperlink Induced Topic Search) – один из двух самых популярных на сегодня в области информационного поиска.

Алгоритм HITS обеспечивает выбор из информационного потока лучших «авторов» (первоисточников) и «посредников» (документов от которых идут ссылки цитирования). Если ввести понятие матрицы инциденций A, элемент которой a_{ij} равен единице, если документ D_i содержит ссылку на документ D_j, и нулю в противном случае, то алгоритм HITS обеспечивает выбор наиболее авторитетных документов, которые предположительно соответствуют собственным векторам матриц AA^T и A^TA с наибольшими модулями собственных значений. В этом смысле алгоритм HITS эквивалентен LSI. Действительно, пусть, в соответствии с сингулярным разложением $A= USV^T$, S – квадратная диагональная матрица. Тогда $AA^T = USV^T\ VSU^T = US\ I\ SU^T = US^2U^T$, где S^2 – диагональная матрица с элементами s_{ii}^2. Очевидно, как и при LSI, собственные векторы, соответствующие наибольшим сингулярным значениям AA^T и/или A^TA будут соответствовать наиболее статистически важным авторам и/или посредникам.

Метод LSI является, пожалуй, одним из самых строгих методов кластерного анализа, однако, неизвестны его непосредственные применения для индексирования динамических информационных потоков (например, в рамках пространственно-векторной модели), по-видимому, ввиду необходимости постоянного пересчета матриц сингулярного разложения при добавлении новых элементов [6]. Кроме того, ввиду своей вычислительной трудоемкости (равной $O(N^2)$, N – размерность A), этот метод применяется только для относительно небольших матриц. Вместе с тем, для решения задач, аналогичных приведенной выше, он вполне пригоден.

4. Взаимосвязь рубрик и метод k-means

Остановимся на еще одном методе кластеризации. Рассмотрим множество уникальных термов из всех информационных портретов: $W = \{w`_k\}$ (k=1,..,N), где N – количество уникальных термов. Рассмотрим проекцию множества информационных портретов $P_i$ на W: $P` = \{ p`_{ij} \}$ (i=1,K. j=1,..,N), где $p`_{ij} = 1$, если $\exists$ k: $w`_k = p_{ij}$, в противном случае $p`_{ij} = 0$.

Тогда произведение матриц $E = P`^T * P`$ будет таблицей взаимосвязей рубрик, построенной в результате анализа состава термов соответствующих информационных портретов.

Естественно, близость двух рубрик i и j в такой модели выражается элементом e_{ij}.

Укрупнение рубрик – актуальная задача кластерного анализа и она может быть решена путем их группировки по признакам подобия и перенумерации (преобразования матрицы E – одновременной перестановки строк и столбцов).

Укрупнение рубрик позволяет, например, эффективней группировать документы потока в предметно-ориентированные базы данных, автоматически формировать и реализовывать наглядную визуализацию классов документов на Web-сайтах и т.д.

Покажем, как можно выделить некоторое число групп взаимосвязанных рубрик методом кластерного анализа k-means [7].

Алгоритм k-means - самый известный из алгоритмов кластеризации, при этом существуют две реализации, «жесткая», когда число k фиксировано и «мягкая», которая позволяет на основании некоторых критериев оценить значение k. Остановимся подробнее на первой реализации.

Рассмотрим векторы-строки матрицы E – E_i (очевидно, ввиду симметричности матрицы E можно было бы рассматривать и столбцы). Простая задача оптимальной группировки векторов E_i в данном случае усложняется необходимостью при перестановке (перенумерации) номеров векторов-строк одновременно переставлять соответствующие их компоненты для сохранения симметрии матрицы E.

Суть алгоритма k-means определяется следующим образом: случайным образом выбирается k векторов-строк, которые определяются как центроиды (наиболее типичные представители) кластеров. Затем k кластеров наполняются – для каждого из оставшихся векторов-строк определяется близость к центроиду соответствующего кластера. После этого вектор-строка приписывается к тому кластеру, к которому он наиболее близок. После этого строки-векторы группируются и перенумеровываются в соответствии с полученной группировки. Затем для каждого из новых кластеров заново определяется центроид – вектор-строка, наиболее близкая ко всем векторам из данного кластера (например,

тот, сумма скалярных произведений которого с каждым из векторов кластера - минимальна). После этого заново выполняется процесс наполнения кластеров, затем вычисление новых центроидов и т.д., пока процесс формирования кластеров не стабилизируется (или набор центроидов не повторится).

Ниже приведен формальный алгоритм k-means:

```
Произвольній выбор центроидов k-кластеров
while процесс формирования не стабилизировался do
    for каждого вектора-строки do
        найти кластер с, центроид которого наиболее близок к вектору-
строке
                                    приписать вектор-строку кластеру с
    end for
    for каждого кластера с do
        вычисление центроида кластера по входящим в него элементам
    end for
    for каждого вектора-строки do
        переставить элементы в векторе-строке,
                                    соответствующие выполненной
перенумерования
    end for
  end while
```

На рис. 1. показана таблица взаимосвязей рубрик общей тематики, которая строится в системе InfoStream® [9]. В таблице явно выделено два кластера – блоков у краев диагонали. В данном случае, верхний кластер соответствует социокультурным аспектам (религия, культура, образование), а нижний – технологическим (связь, компьютеры, наука и техника).

В отличие от метода LSI, k-means идеально подходит для кластеризации динамических информационных потоков. Если вернуться к рассмотрению пространственно-векторной модели, то на момент t можно выделить K тематических кластеров из информационного потока D(t), а каждый документ рассматривается как множество термов. Мера близости между отдельными документами может определяться метрикой μ как скалярное произведение соответствующих этим документам векторов, элементами которых являются те же значения TF*IDF.

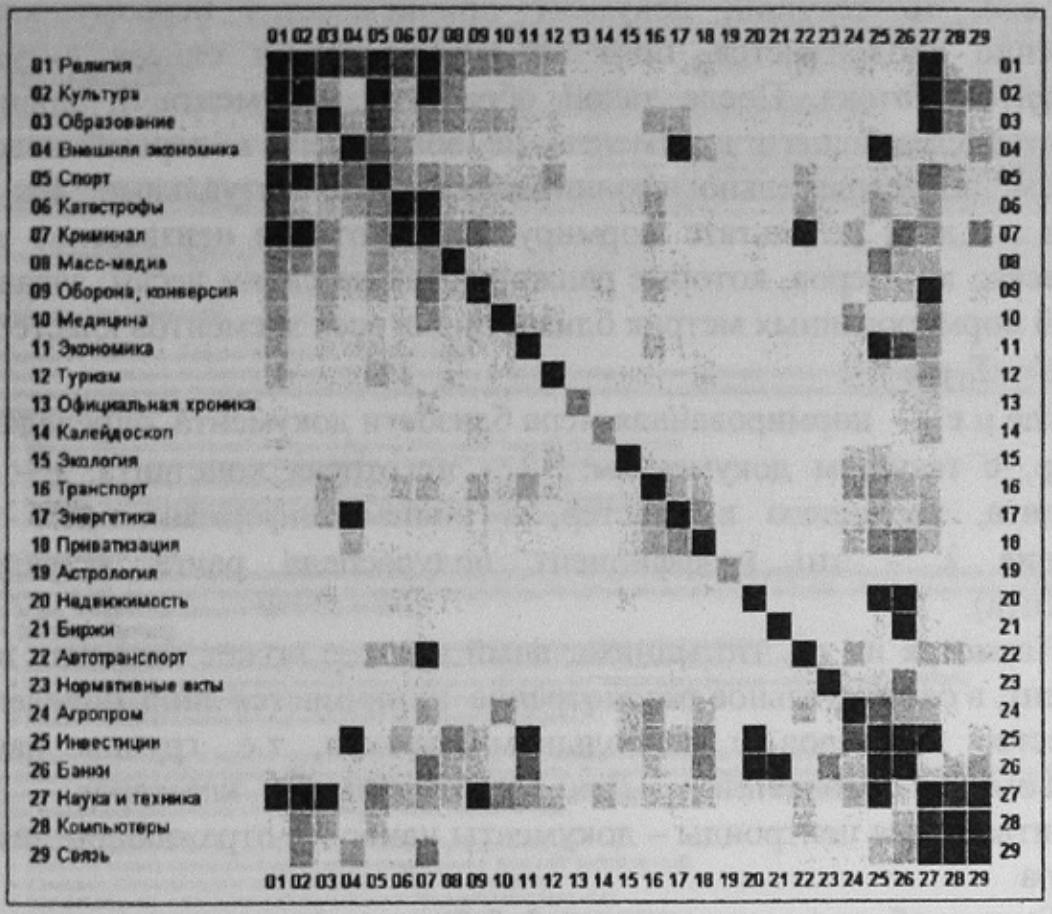

Рис.1. Таблица взаимосвязей рубрик

5. Выявление сюжетов из информационных потоков

Если к потоку документов за некоторый промежуток время Δt добавляется несколько новых документов, то как и в приведенном выше алгоритме, каждый из новых документов приписывается к соответствующему кластеру в соответствии с мерой μ, после чего происходит пересчет центроидов. Рекурсивный пересчет наполнения кластеров в соответствии с новыми центроидами может выполняться с заданной заранее периодичностью.

С помощью некоторой модификации этого метода можно, например, достаточно эффективно строить цепочки основных сюжетов, формируемые в информационном документальном потоке. Например, алгоритм выявления основных сюжетных цепочек, используемый в системе InfoStream [10] заключается в следующем. Последний поступивший на вход системы документ (документ с номером 1 при обратной нумерации) порождает первый кластер и сравнивается со всеми предыдущими в соответствии с приведенной выше метрикой μ. Если эта мера близости для какого-нибудь документа оказывается ближе заданной

пороговой, то текущий документ приписывается первому кластеру. Сравнение продолжается, пока не исчерпывается список актуальных документов потока. После такой обработки документа 1, происходит обработка следующего документа, не вошедшего в первый кластер, с которым последовательно сравниваются все актуальные документы потока и т.д. В результате формируется некоторое неизвестное заранее количество кластеров, которые ранжируются по своим весам, задаваемым суммой нормированных метрик близости для всех элементов кластера:

$C = \Sigma_j \mu e^{-\lambda j}$

где $\mu e^{-\lambda j}$ - нормированная мера близости документа, определяющего кластер, с текущим документам, λ - некоторая константа, j – номер документа, входящего в кластер, в общем информационном потоке (значение λ - это коэффициент полураспада ранга актуальности документа).

Несмотря на то, что минимальный кластер может включать всего 1 документ, в окончательное рассмотрение принимается лишь определенное количество кластеров с наибольшими весами, т.е. группы наиболее актуальных сообщений. Для выбранных кластеров заново пересчитываются центроиды – документы наиболее отражающие тематику кластера.

Таким образом, в системе InfoStream формируются сюжетные цепочки, реализующие запросы типа «о чем пишут больше всего в последнее время?» (Рис.2).

6. Выводы: Преимущество подхода «два в одном»

Отметим одно преимущество рубрицирования путем взвешивания потоков документов в пространстве информационных портретов тематических рубрик.

Чаще всего информационные портреты формируются путем лингво-статистического анализа массивов документов, полученных в результате поиска по соответствующим тематическим запросам. Эти запросы в большинстве промышленных информационно-поисковых системах составляются экспертами на языках, являющихся расширением булевой алгебры операторами контекстной близости.

Окончательная же рубрикация документов в пространстве ИП предполагает более «экономный» весовой подход на основе массива термов, полученных в результате периодической отработки булевых запросов.

Таким образом, в результате учитываются «логические» преимущества первого подхода и эксплуатационные – второго.

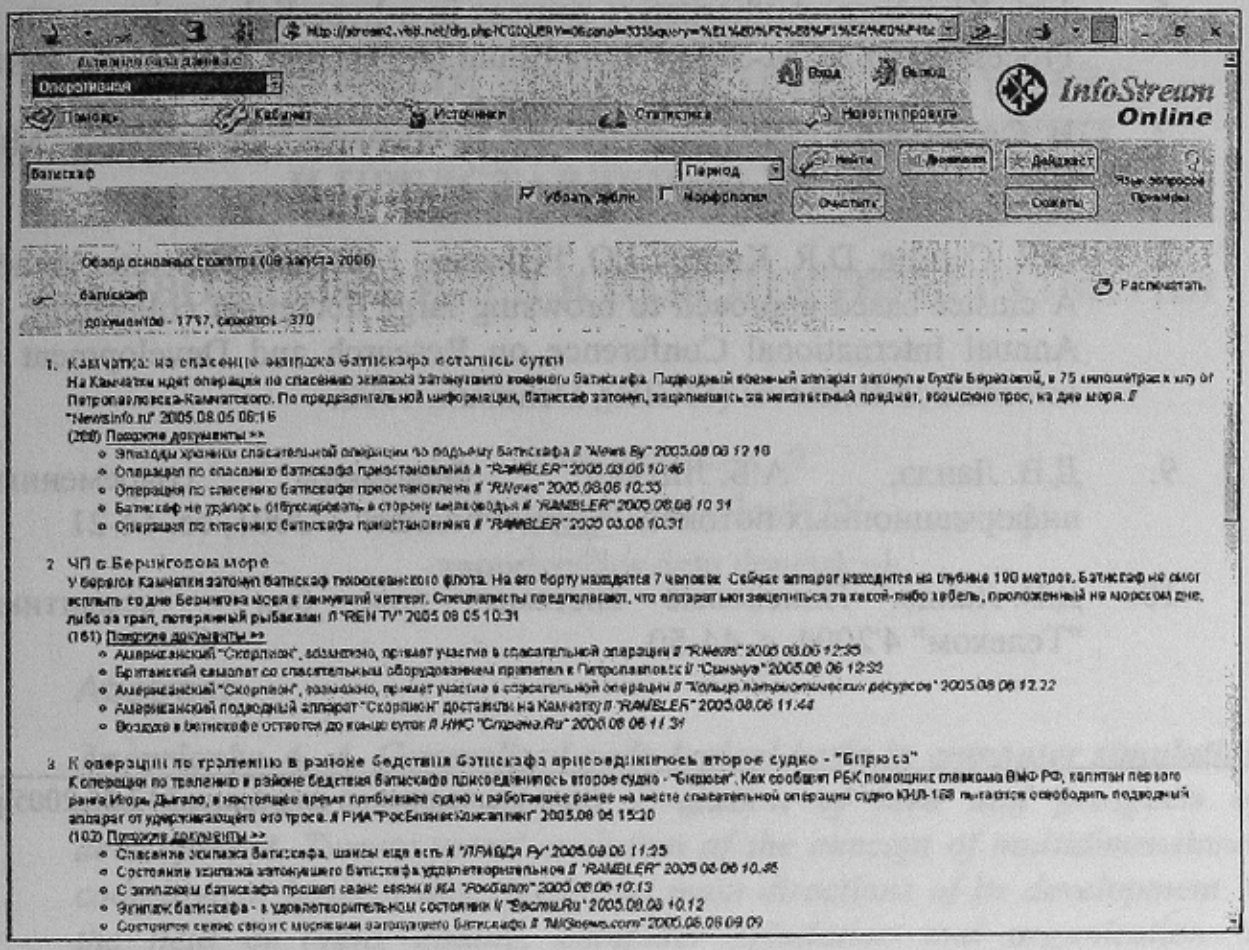

Рис.2. Пример формирования сюжетных цепочек

Литература

1. Д.В. Ландэ. Поиск знаний в Internet. Профессиональная работа. СПб: Диалектика/Вильямс, 2005 г. 272 с.

2. А.В. Антонов. Методы классификации и технология Галактика-Зум. НТИ. Серия 1. Выпуск 6. 2004 год. сс. 20-27

3. Chakrabarti Soumen, Mining the web. Discovery knowledge from hypertext data// Publisher: Morgan Kaufmann, 2002. 344 p.

4. И.А. Титов, Дж. Хендерсон. Метод синтаксического разбора с использованием определяемых обучающим набором ядер, построенных на основе вероятностных моделей. Труды Международного семинара «Диалог'2005».

5. G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer, R. A. Harshman, L.A. Streeter, and K.E. Lochbaum. Information retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. ACM SIGIR, 1988

6. J.M. Kleinberg. Authoritative sources in a hyperlink environment. In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998.

7. И. Сегалович. Как работают поисковые системы, “Мир Интернет” 10’2002

8. D.R. Cutting, D.R. Karger, J.O. Pedersen, J.W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Annual International Conference on Research and Development in Information Retrieval (SIGIR), Denmark, 1992.

9. Д.В. Ландэ, А.Б. Литвин. Феномены современных информационных потоков. "Сети и бизнес" 1’2001, сс. 14-21

10. Д.В. Ландэ. Поисковые системы: поле боя – семантика, "Телеком" 4'2004, с. 44-50.

Дата надходження до редакції 12.06.2005 р.