УДК 681.3

Д. В. Ландэ
Информационный центр «ЭЛВИСТИ»
ул. М. Кривоноса, 2а, 03037 Киев, Украина

Подход к анализу новостных потоков как дискретных сигналов

Описана модель, в которой текстовые информационные потоки рассматриваются как дискретные сигналы, в качестве амплитудных значений которых выступают частотно-семантические ранги наиболее рейтинговых терминов или документов. Обоснован подход к созданию инструментария, обеспечивающего просмотр так называемых «маргинальных» сообщений по тематике, определяемой запросом пользователя, то есть фактически дающего ответ на вопрос, о чем пишут меньше всего в рамках данной тематики в последнее время.

Ключевые слова: *информационные потоки, обработка сигналов, Интернет, текстовый корпус, ранжирование*

Исследование новостной составляющей информационного пространства Интернет, то есть потока новостных сообщений, публикуемых на страницах web-сайтов, должно использовать принципиально новый инструментарий, так как классические методы сегодня уже не всегда приемлемы ввиду резкого увеличения объемов и динамики информационных потоков [1].

Одна из идей, к которой все чаще обращаются в настоящее время, заключается в анализе текстовых массивов как дискретных сигналов, определяемых частотно-семантическими рангами [2] ключевых слов или отдельных сообщений.

В этой статье рассматривается модель, в которой аналогами дискретных сигналов выступают ключевые слова (наиболее ранговые термины) из сообщений, или отдельные сообщения информационных потоков, порождаемых информационными web-сайтами. В соответствии с приведенным ниже алгоритмом каждому сообщению приписывается вес, который равен усредненной частоте появления во всем информационном потоке входящих в это сообщение значимых ключевых слов. Очевидно, чем меньше этот вес, тем документ более уникален.

Понятно, что для информационного наполнения модели необходимо использовать достаточно мощный текстовый корпус, который был доступен автору — это база данных системы контент-мониторинга InfoStream [3]. Система InfoStream применяется для решения задач автоматизированного сбора новостной информа-

ции с открытых web-сайтов, а также обеспечения доступа к ней в поисковых режимах. Эта разработанная в компании ElVisti система в настоящее время охватывает ретроспективные базы данных, представляющие собой текстовый корпус объемом свыше 20 млн. документов из 2000 источников информации.

Обработка входных сообщений в системе контент-мониторинга InfoStream и поступление их в рассматриваемую аналитическую модель выполнялась по следующей схеме.

Новостные сообщения → конвертирование в формат системы (в том числе автоматическая рубрикация) → стемминг (морфологическая обработка, усечение флексий) → выделение ключевых слов (в рассматриваемой модели до 12) → аналитическая модель.

Ниже приведен двухпроходный алгоритм формирования словаря уникальных слов из входного массива из N сообщений, а затем вычисления весов отдельных сообщений.

Этап 1: первичная обработка входного информационного массива

```
while количество необработанных сообщений из массива > 0 do
    чтение текущего сообщения
    for каждого сообщения do
        while не исчерпался список ключевых слов do
            for каждого ключевого слова do
                if ключевое слово уже входит в словарь
                then вес ключевого слова = вес ключевого слова + 1
                else добавить ключевое слово в словарь с весом 1
            end for
        end while
    end for
end while
```

Этап 2: повторная обработка информационного массива

```
while количество необработанных сообщений из массива > 0 do
    чтение текущего сообщения
    вес сообщения = 0
    for каждого сообщения do
        счетчик ключевых слов = 0
        while не исчерпался список ключевых слов do
            for каждого ключевого слова do
                определение веса из словаря уникальных слов
                вес сообщения = вес сообщения + вес слова
                счетчик ключевых слов = счетчик ключевых слов + 1
            end for
        end while
    end for
    вес сообщения = вес сообщения / число ключевых слов
end while
```

Таким образом, вес сообщения определяется по формуле:

$$W_D = \frac{\sum_{w \in D} w}{|D|},$$

где W_D — вес сообщения; w — ключевое слово из сообщения; $|D|$ — количество ключевых слов в документе (в рассматриваемой модели $1 \le |D| \le 12$). Как видно, при значениях β в указанном выше диапазоне w является монотонно возрастающей функцией от n.

Как следует из алгоритма, каждое сообщение в данной модели рассматривается как массив ключевых слов (Bag of Words [4]), хотя при построении модели учитывались структурные особенности сообщений [5], в частности, при определении веса ключевых слов учет их местоположения в тексте.

В классической пространственно-векторной модели [6] значения рангов отдельных ключевых слов определяется формулой *TF·IDF*. В данном случае *TF* — это локальная частота ключевого слова (Term Frequency), а *IDF* — величина, обратная частоте встречаемости во всем потоке документов, содержащих данный терм (Inverse Document Frequency).

В то время как локальная частота ключевого слова в документе говорит о его значимости в пределах документа, то обратная частота встречаемости свидетельствует об уникальности ключевого слова во всем потоке документов.

В рассматриваемой модели в соотношении *TF·IDF* фактически анализируется лишь второй сомножитель (а точнее, обратная ему величина), исходя из того, что заведомо высокий уровень значений *TF* определяется процедурой выявления ключевых слов, выполняемой ранее системой контент-мониторинга.

В рамках модели в качестве веса ключевых слов используется частота их появлений во входном информационном потоке. В свою очередь, эта частота зависит от объема самого потока и от количества уникальных слов, то есть объема автоматически формируемого словаря уникальных слов. В компьютерной лингвистике эмпирический закон Хипса [7] связывает объем документа с объемом словаря уникальных слов, входящих в этот документ. В соответствии с законом Хипса, эти значения связываются соотношением:

$$v(n) = Kn^{\beta},$$

где v — объем словаря уникальных слов, составленный из текста, состоящего из n уникальных слов; K и β — определяемые эмпирически параметры. Для европейских языков K принимает значения от 10 до 100, а β — от 0,4 до 0,6.

В случае анализа не полных текстов, а фиксированного количества нормированных ключевых слов, эти параметры изменяются, однако сама закономерность Хипса остается в силе (рис. 1).

Джордж Зипф [8] экспериментально показал, что, если для какого-либо достаточно большого текста составить список всех встретившихся в нем слов, а затем ранжировать эти слова в порядке убывания частоты встречаемости в тексте, то для любого слова произведение его ранга в этом списке и частоты встречаемости в тексте будет величиной постоянной, то есть $f \cdot r = c$, где f — частота встречаемо-

сти слова в тексте; r — ранг слова в списке; c — эмпирически определяемая константа.

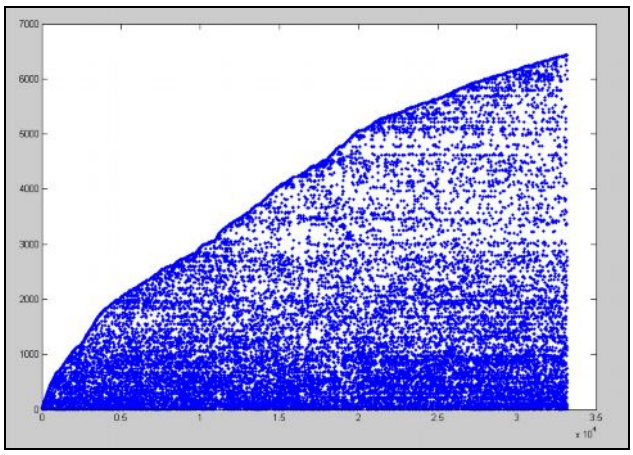


Рис. 1. График зависимости количества уникальных ключевых слов от общего количества ключевых слов потока подчиняется закону Хипса. При этом $K = 4$, $\beta = 0{,}65$

В рассматриваемой же нами модели в соответствии с приведенным выше алгоритмом распределение весов ключевых слов вполне вписывается в закон Зипфа (рис. 2), сформулированный изначально для ранговых распределений ненормированных слов в полнотекстовых документах. Однако в модели вместо ранжированного сортированного словаря используется простой порядковый номер. Феномен объясняется тем, что в соответствии с положениями математической статистики большая часть наиболее часто встречающихся слов попадает в некоторое ограниченное количество первых по порядку сообщений.

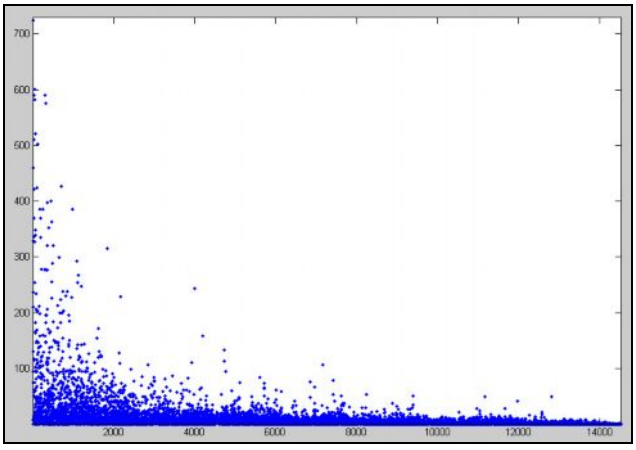


Рис. 2. Зависимость частоты уникальных слов в потоке от их порядковых номеров

Статистически связанная с названными выше закономерностями зависимость параметров распределения весов отдельных сообщений от их порядковых номеров в потоке (рис. 3) имеет вполне определенное смысловое объяснение. Оказы-

вается, что амплитуда этого распределения возрастает с увеличением количества сообщений в потоке (рис. 4). Действительно, средний вес уникального ключевого слова равен общему числу слов из потока, разделенному на количество уникальных слов:

$$w(n) = n/v(n) = n^{1-\beta} / K.$$

Этому же значению равно и математическое ожидание веса отдельного сообщения из потока.

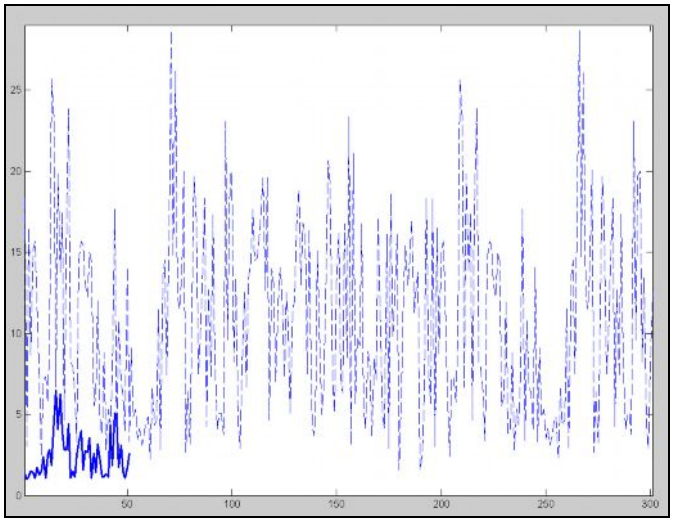


Рис. 3. Графики зависимости веса сообщений от их номеров в потоке. Рассматривается два информационных потока (50 и 300 сообщений)

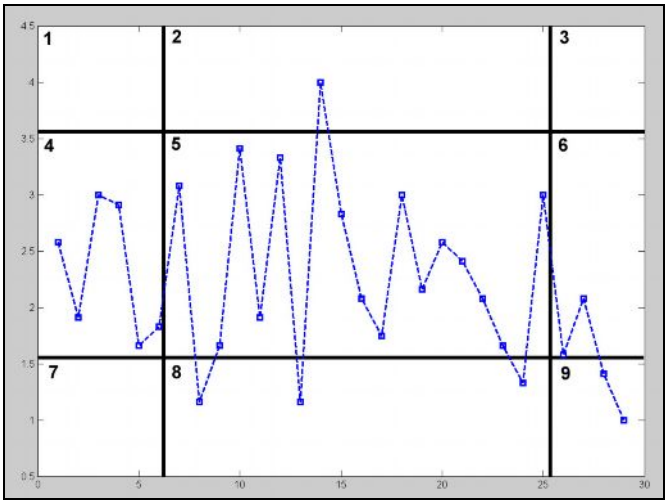
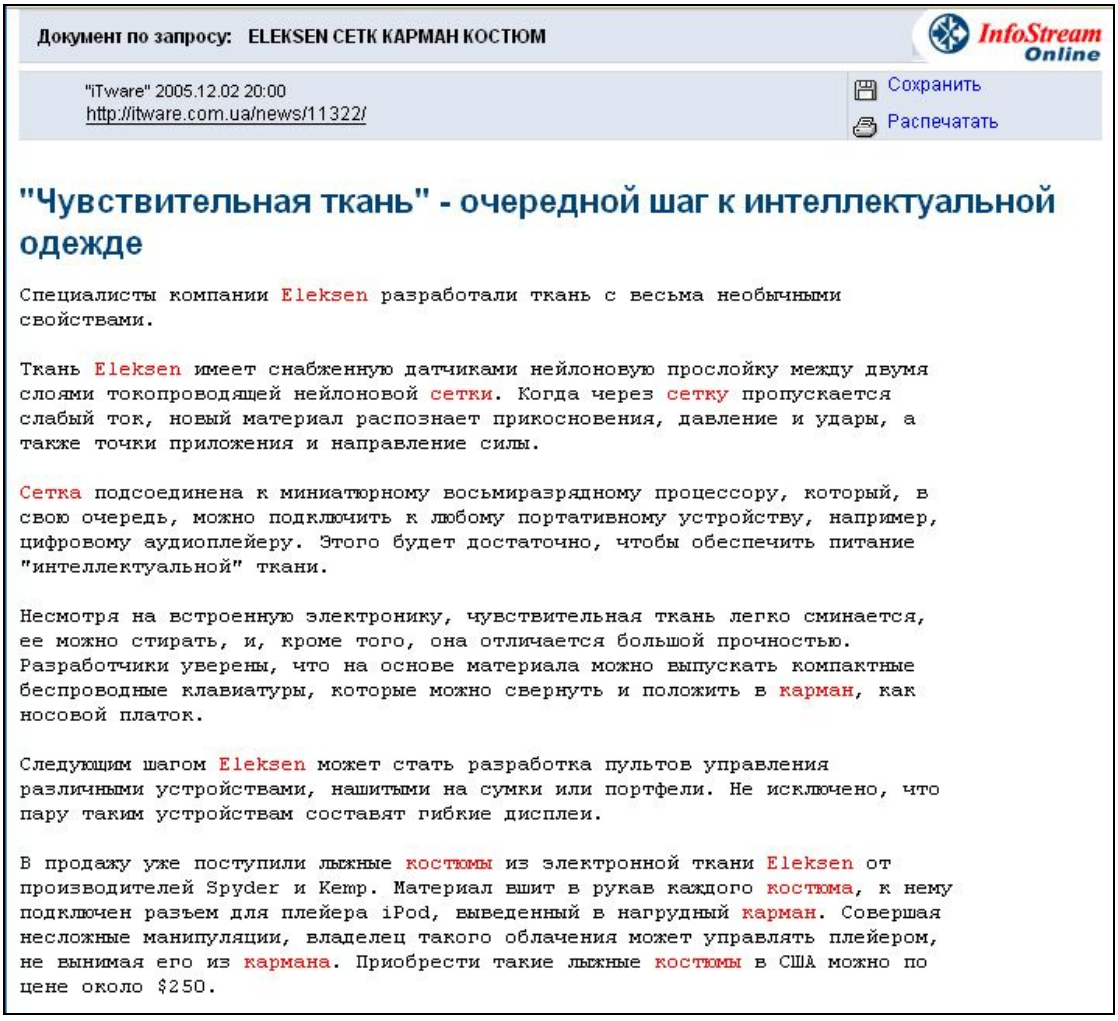


Рис. 4. Основные области графика распределения весов сообщений

Изображенные на рис. 4. основные области графика дискретного сигнала, соответствующего информационному потоку, можно охарактеризовать следующим

образом. Горизонтальные зоны: 1, 2, 3 — топ-новости; 4, 5, 6 — мэйнстрим; 7, 8, 9 — маргинальная зона. Вертикальные зоны: 1, 4, 7 — устаревающие сообщения; 2, 5, 8 — основная тематика; 3, 6, 9 — последние известия.

На рис. 5 приведен документ, попавший в маргинальную зону при анализе потоков сообщений по компьютерной тематике, полученных с web-сайта ITWARE (http://itware.com.ua). Этот пример с очевидностью подтверждает уникальность содержания сообщений из этой области по сравнению с мэйнстрим-сообщениями по информационным технологиям. Это всего лишь одно из многих практических подтверждений корректности данной модели, подхода к созданию инструментария в рамках системы контент-мониторинга, обеспечивающего просмотр маргинальных сообщений по тематике, определяемой запросом, то есть фактически дающего ответ на вопрос, о чем пишут меньше всего в рамках данной тематики в последнее время. Этот инструментарий логически дополняет уже существующий в системе InfoStream сервис получения сюжетов из наиболее популярных сообщений [9].



Документ по запросу: ELEKSEN СЕТК КАРМАН КОСТЮМ

InfoStream Online

"iTware" 2005.12.02 20:00
http://itware.com.ua/news/11322/

Сохранить
Распечатать

"Чувствительная ткань" - очередной шаг к интеллектуальной одежде

Специалисты компании Eleksen разработали ткань с весьма необычными свойствами.

Ткань Eleksen имеет снабженную датчиками нейлоновую прослойку между двумя слоями токопроводящей нейлоновой сетки. Когда через сетку пропускается слабый ток, новый материал распознает прикосновения, давление и удары, а также точки приложения и направление силы.

Сетка подсоединена к миниатюрному восьмиразрядному процессору, который, в свою очередь, можно подключить к любому портативному устройству, например, цифровому аудиоплейеру. Этого будет достаточно, чтобы обеспечить питание "интеллектуальной" ткани.

Несмотря на встроенную электронику, чувствительная ткань легко сминается, ее можно стирать, и, кроме того, она отличается большой прочностью. Разработчики уверены, что на основе материала можно выпускать компактные беспроводные клавиатуры, которые можно свернуть и положить в карман, как носовой платок.

Следующим шагом Eleksen может стать разработка пультов управления различными устройствами, нашитыми на сумки или портфели. Не исключено, что пару таким устройствам составят гибкие дисплеи.

В продажу уже поступили лыжные костюмы из электронной ткани Eleksen от производителей Spyder и Kemp. Материал вшит в рукав каждого костюма, к нему подключен разъем для плейера iPod, выведенный в нагрудный карман. Совершая несложные манипуляции, владелец такого облачения может управлять плейером, не вынимая его из кармана. Приобрести такие лыжные костюмы в США можно по цене около $250.

Рис. 5. Сообщение по компьютерной тематике из маргинальной зоны

В заключение заметим, что предложенная модель охватывает лишь некоторые частотно-семантические подходы к рассмотрению текстовых информационных потоков как дискретных сигналов. Получены первые результаты исследования, которое может включать в себя более полный учет структурных особенностей текстов, анализ корреляции сигналов, фильтрацию типа «сигнал–шум» и т.д. Можно также предположить, что к обработке текстовых потоков будут применимы такие популярные сегодня техники обработки сигналов как анализ главных компонент, слепое разделение источников, вейвлеты.

1. *Ландэ Д.В., Брайчевский С.М.* Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1. — 2005. — № 11. — С. 21–33.

2. *Del Corso G.M., Gulli A., and Romani F.* Ranking a Stream of News // Proc. 14-th International World Wide Web Conference. — Chiba (Japan). — 2005. — P. 97–106.

3. *Ландэ Д.В.* Сканер системы контент-мониторинга InfoStream // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. — Харьков: НАКУ «ХАИ», 2005. — Вып. 28. — С. 53–58.

4. *Salton G., Allan J. and Buckley C.* Approaches to Passage Retrieval in Full Text Information Systems // ACM SIGIR Conference on R&D in Information Retrieval. — 1993. — P. 49–58.

5. *Baeza-Yates R. and Ribeiro-Neto B.* Modern Information Retrieval. — Addison-Wesley, 1999.

6. *Chakrabarti Soumen.* Mining the Web. Discovery Knowledge from Hypertext Data. — San Francisco: Publisher Morgan Kaufmann, 2002. — 344 p.

7. *Heaps H.S.* Information Retrieval: Computation and Theoretical Aspects. — Orlando: Academic Press Inc., FL, 1978. — P. 206–208.

8. *Zipf, George Kingsley*. Human Behaviour and the Principle of Least Effort. — Cambridge: Wesley, MA, 1949.

9. *Ландэ Д.В.* Поиск знаний в Internet. Профессиональная работа. — М.: Вильямс, 2005. — 272 с.

Поступила в редакцию 02.02.2006