СПИСОК ЛИТЕРАТУРЫ

1. Пупков К. А., Коньков В. Г. Интеллектуальные системы.— М.: изд-во МГТУ им. Н. Э. Баумана, 2003.— 348 с.

2. Гришин Ю П., Казаринов Ю. М. Динамические системы, устойчивые к отказам.— М.: Радио и связь, 1985.— 176 с.

3. Романов В. П. Интеллектуальные информационные системы в экономике: Учеб. пособие / Под ред Н. П. Тихомирова.— М.: Экзамен, 2003.— 496 с.

4. Петров В. Н. Информационные системы. - СПб.: Питер, 2003 — 688 с.

5. Воинов Б. С. Информационные технологии и системы: В 2-х кн. Кн. 2. Прикладные системные исследования.- Нижний Новгород: изд-во ННГУ им. Н. И. Лобачевского, 2001.— 272 с.

6. Калман Р., Фалб П., Арбиб М. Очерки по математической теории систем / Пер. с англ.-- М.: Мир, 1971.— 400 с.

7. Муромцев Ю. Л., Ляпин Л. Н., Грошев В. Н., Шамкин В. Н. Теоретические основы исследования сложных систем с учетом надежности: Учеб. пособие / Моск. ин-т хим. машиностроения.— М.: изд-во МИХМ, 1987.— 116 с.

8. Муромцев Ю. Л., Ляпин Л. Н., Попова О. В. Моделирование и оптимизация технических систем при изменении состояний функционирования.- Воронеж: ВГУ, 1992.— 164 с.

9. Аоки М. Оптимизация стохастических систем. - М.: Наука, 1971.— 424 с.

10. Kikuchi N. Control problems of contingent equation // Publ. RIMS Kyoto Univ. Ser. A. 1969. - Vol. 3, № 1.— P. 58 99.

11. Благодатских В. И. Некоторые результаты по теории дифференциальных включений // Summer school on Ordinary Differential Equations. Par 2. - Brno, 1974.-- p. 29-67.

12. Филипов А. Ф. Дифференциальные уравнения с разрывной правой частью.- М.: Наука, 1985.— 224 с.

13. Артемьев В. В. Теория динамических систем со случайными изменениями структуры. Минск: Высш. шк., 1979.- 160 с.

14. Вонэм В. М. Стохастические дифференциальные уравнения в теории управления // Математика: Сб. пер. - 1973. - Т. 17: № 4 С. 82 114.

15. Казаков Е. И., Артемьев В. М. Оптимизация динамических систем случайной структуры. - М.. Наука, 1980. - 384 с.

16. Забелло Л. Е. К теории управляемости линейных нестационарных систем // Дифференциал уравнения.- 1973. Т. 9, № 3. С. 563 564.

17. Мишулина О. А. Анализ линейных систем управления со случайным скачкообразным изменением параметров с применением ЭВМ // Изв. АП СССР. Техн. кибернетика. 1968. № 3. С. 128 134.

18. Воронов А. А. Введение в динамику сложных управляемых систем. М.: Наука, 1985. 352 с.

19. Муромцев Ю. Л. Определение границ показателей надежности сложных систем // Автоматика и телемеханика. 1977. № 9. С. 177 190.

Материал поступил в редакцию 27.07.06.

УДК 004

Д. В. Ландэ, В. Н. Фурашев

Выявление новых событий в рамках системы контент-мониторинга

Освещается популярная в настоящее время во всем мире тематика выявления, отслеживания и группировки новых событий из потоков новостей (New Event Detection, Tracking, Clustering). Приведен краткий обзор теоретических и практических разработок в этом направлении. Представлены оригинальные критерии, используемые для выявления новых событий в рамках системы контент-мониторинга InfoStream®.

ВВЕДЕНИЕ

Современный мир характеризуется чрезвычайно высокой динамикой изменения ситуаций. На эту динамику влияют, помимо всего, и процессы развития демократичных начал в организации и управлении обществом, которые наблюдаются в настоящее время практически во всех странах мира, особенно в странах так называемого "постсоветского пространства". На принятие решений любого плана (политических, экономических, социальных и т. д.) и на любом уровне (личностном, региональном, государственном, межгосударственном и т. д.) несомненное влияние оказывает не только степень информирования в области принятия решения, но и уровень возможной информационной реакции на принятое решение. Если традиционными средствами массовой информации можно хоть как-то управлять административно-экономически, то на информацию, которая распространяется с помощью всемирной компьютерной сети Интернет "управы", практически ни одно государство мира не имеет. Объективности ради заметим, что попытки влиять на информацию, циркулирующую в Интернете, постоянно предпринимались, предпринимаются, и мы вполне уверены, будут предприниматься. Пути этого влияния различны от полного запрета гражданам страны пользоваться возможностями системы Интернет до принудительного мониторинга Интернет-провайдерами использования и направленности информационных ресурсов

определенными категориями пользователей Интернет-пространства. Однако все эти попытки пока не привели "инициаторов" таких ограничений к успехам. Больше того, мы наблюдаем постоянный прирост всевозможной информации в сети Интернет. Эта тенденция не только сохранится, но и будет распространяться на все новые области человеческой деятельности, буквально на всех уровнях. Однако, как известно, любая медаль имеет две стороны. Именно обилие информации помогает принимать более правильные, более взвешенные решения, но это же обилие приводит и к неоперативности принятия необходимых решений в силу увеличения сроков её "переработки и осмысливания".

Необходимо также отметить, что на современном этапе развития общества Интернет-пространство перестало быть экзотикой, уделом только для "избранных". На сегодняшний день сеть Интернет с её бесчисленным количеством информационных Web-сайтов стала неотъемлемым атрибутом информационной поддержки принятия решений на любом уровне.

Именно рост объемов информации и скорость ее распространения фактически породили понятие "информационные потоки" [1]. Исследование такой составляющей этих потоков, как сообщения, публикуемые на страницах Web-сайтов, должно использовать принципиально новый инструментарий, так как классический математический аппарат и инструментальные средства не всегда способны адекватно отражать ситуацию. В этом случае речь идет не столько об анализе документального массива фиксированного размера, сколько о навигации в потоке документов.

Несомненным является тот факт, что информационные потоки большей частью порождаются событиями реального мира, которые необходимо оперативно учитывать при принятии тех или иных решений. Действительно, возникновение информационных потоков можно представить себе как генерацию и движение наборов данных, ассоциированных с определенным событием, реализуемым как некоторый смысловой блок. Конечно, одному событию может соответствовать произвольное число сообщений. Таким образом, характеристики информационных потоков изначально определяются потоками событий реального мира.

Если событие новое и важное, то о нем будут много говорить в дальнейшем, т. е. задача выявления новых событий из потока новостей является задачей предсказания дальнейшего появления множества "подобных" сообщений — задачей прогноза [2].

СИНДИКАЦИЯ НОВОСТЕЙ

Оптимальное решение, способное помочь ориентироваться в динамической части Интернета, сегодня предоставляют системы синдикации новостей [3, 4]. Под синдикацией в данном случае понимается сбор информации в Интернете и последующее распространение ее фрагментов в соответствии с потребностями пользователей.

Технология синдикации Интернет-новостей (рис. 1) включает в себя "обучение" программ сбора информации структурным особенностям отдельных источников (Web-сайтов), непосредственное сканирование информации, приведение ее к общему формату (как правило, XML), а также классификацию.

Средства классификации и распределения информации представляют собой информационно-поисковую систему избирательного распространения информации (Информационный роутер). Документы, поступающие в систему, анализируются на соответствие тематическим запросам. Релевантные документы рассылаются пользователям, а также загружаются в тематические базы данных.

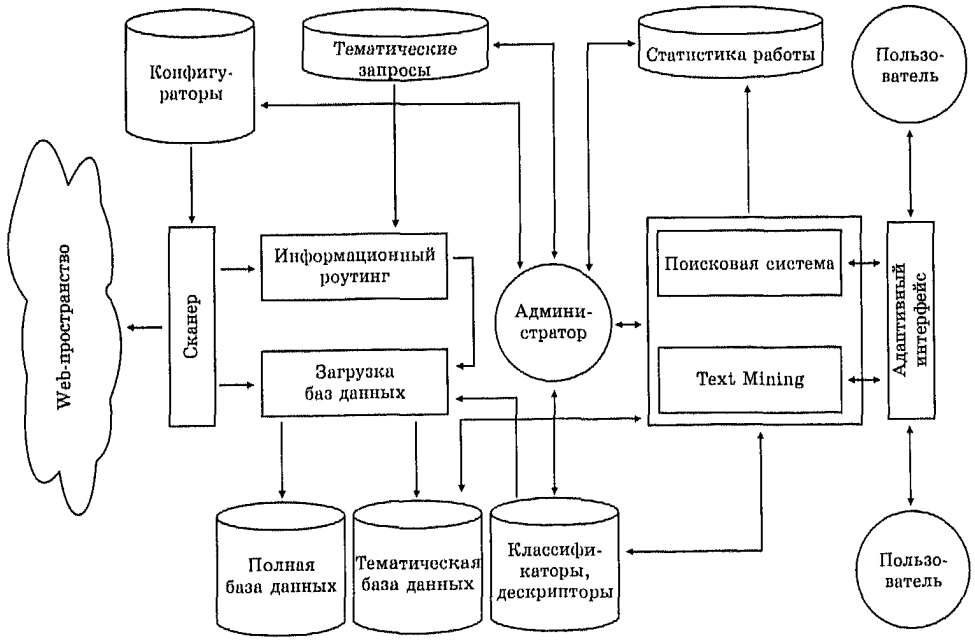

Рис. 1. Функциональная схема системы контент-мониторинга новостей из Интернета

В режиме диалогового доступа к базам данных обеспечивается просмотр, поиск и отображение данных, а также предоставляется возможность обращения к оригиналам документов в Интернет.

Перспективным направлением развития технологии контент-мониторинга является глубинный анализ текстов (Text Mining), средствами которого обеспечивается решение задач формирования тематических информационных каналов, дайджестов, таблиц взаимосвязей и гистограмм распределения понятий. К этому классу средств можно также отнести выявление новых событий, их отслеживание и группировку (кластеризацию).

ОБЩИЙ ПОДХОД К ВЫЯВЛЕНИЮ НОВЫХ СОБЫТИЙ

Задача выявления новых событий из потока сообщений предполагает, что на вход соответствующего программно-технологического комплекса последовательно поступают новые документы как непосредственно от средств сканирования (политематический поток), так и от информационного роутера — системы избирательного распространения информации, — отобранные по тематическому запросу (рис. 2). Далее в соответствии с определенными алгоритмами (некоторые из них приведены ниже) происходит выявление новых событий. Новые события описываются в документах, для которых с помощью отдельных программных модулей во временной ретроспективе формируются цепочки подобных документов (сюжетные цепочки). Документы, отражающие различные новые события, могут быть основой новых групп взаимосвязанных документов (кластеров), которые предположительно заполнятся в дальнейшем. В этом предположении и заключается прогнозный момент технологии выявления новых событий. Со временем каждый из кластеров может стать основой формирования полноценной сюжетной цепочки.

Алгоритм выявления основных сюжетных цепочек, используемый, например, в системе контент-мониторинга InfoStream [5], заключается в следующем:

последний поступивший на вход системы документ (документ с номером 1 при обратной нумерации) порождает первый кластер и сравнивается со всеми предыдущими. Если 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 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.

Несмотря на то что минимальный кластер может включать всего один документ, в окончательное рассмотрение принимается лишь определенное количество кластеров с наибольшими весами, т. е. группы наиболее цитируемых и актуальных сообщений. Для выбранных кластеров заново пересчитываются центроиды — документы, наиболее отражающие тематику кластера. Таким образом и формируются сюжетные цепочки, реализующие запросы типа "о чем пишут больше всего в последнее время?"

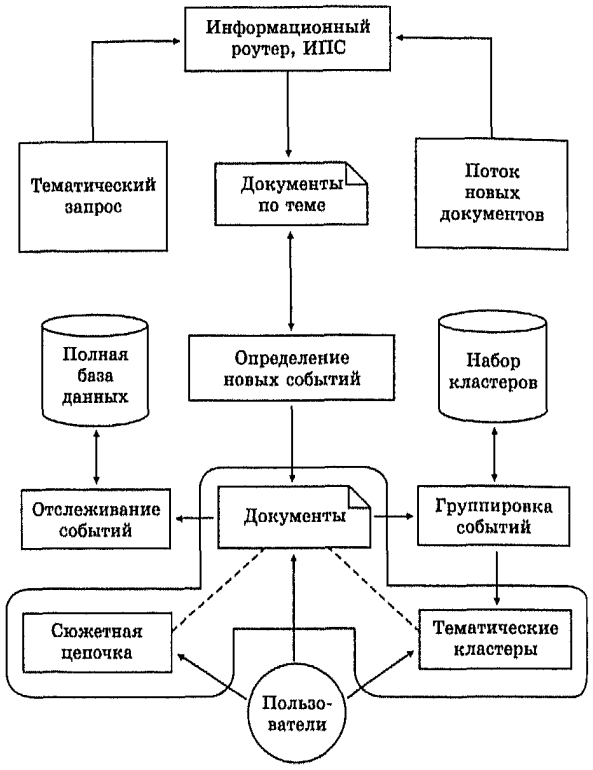

Рис. 2. Определение новых событий — одна из функций системы контент-мониторинга

При построении сюжетных цепочек система определяет лингво-статистические характеристики отобранных в результате поиска документов и автоматически выявляет наиболее значимые темы, освещаемые в информационных потоках. Все весомые сообщения группируются по принадлежности к автоматически определяемым сюжетам. В качестве названий сюжетных цепочек используются заголовки сообщений, наиболее точно отражающих их суть. Порядок отображения сюжетов определяется количеством сообщений в сюжетной цепочке, что отражает общий интерес к данной теме, и временем публикации сообщений.

Сюжетная цепочка выстраивается в результате обработки пользовательского запроса, процесс составления которого в этом случае максимально упрощается: для получения точных результатов вполне достаточно указать одно-два слова, относящихся к необходимой тематике.

Вместе с тем прогнозный вопрос, состоящий в том, что о событии пишут пока мало, но оно достаточно важное и в дальнейшем получит большой резонанс, остается открытым. Этот вопрос связан с общей задачей нахождения исключений или аномалий, т. е. объектов, которые своими характеристиками сильно выделяются из общей массы (хотя в дальнейшем могут породить множество себе подобных).

СУЩЕСТВУЮЩИЕ ПОДХОДЫ К ВЫЯВЛЕНИЮ НОВЫХ СОБЫТИЙ

Подход Солтона (термы)

Подход Солтона [6, 7] заключается в использовании векторно-пространственного представления документов и традиционных методов кластеризации. При этом малый вес приписывается высокочастотным словам из массива документов, что вполне укладывается в модель TF*IDF. Напомним, TF это локальная частота терма (Term Frequency), а IDF — величина, обратная частоте встречаемости во всем потоке документов, содержащих данный терм (Inverse Document Frequency).

Если локальная частота терма в документе говорит о его значимости 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.

Документы при этом подходе обрабатываются последовательно в соответствии с таким алгоритмом:

1. Первому рассматриваемому документу ставится в соответствие первый кластер. Каждый кластер представляется вектором термов (ключевых слов), входящих в документы этого кластера. Нормированный каким-то образом вектор термов принято называть центроидом. Иногда центроидом называют документ, самый близкий по некоторому критерию к вектору термов данного кластера, что не меняет сути данного алгоритма.

2. Каждый следующий документ сравнивается с центроидами существующих кластеров (для этого вводится некоторая мера близости).

3. Если документ достаточно близок к некоторому кластеру, то он приписывается этому кластеру, после чего происходит пересчет соответствующего центроида.

4. Если документ не близок к существующим кластерам, то происходит формирование нового кластера, которому приписывается данный документ.

5. Временной диапазон рассматриваемых документов принято называть "окном наблюдения". Кластеры, все документы которых выходят за пределы окна наблюдения, выносятся за рамки рассмотрения.

В результате работы алгоритма каждому новому возникающему кластеру соответствует новое событие, отражаемое в документах данного кластера.

Подход Папка (запросы)

В соответствии с подходом, предлагаемым Р. Папком [8], новые события выявляются из документов, не удовлетворяющих запросам пользователей, составленным с учетом уже известных событий. Алгоритм выявления новых событий заключается в следующем:

1. Формируются запросы по известным темам (при этом используются техники Text Mining выявления и выбора понятий из текстов сообщений).

2. Новый поступающий документ сравнивается с существующими запросами.

3. Если документ не соответствует запросам, то он ассоциируется с новым событием.

4. В систему включается новый запрос, соответствующий данному документу (опционально).

Дополнительно к приведенному алгоритму подход Папка подразумевает использование механизмов ранжирования результатов поиска (для выбора центроидов кластеров), выявления и выбора понятий [9], а также архитектуру системы избирательного распространения информации InRoute [10].

Многопараметрический подход в рамках системы InfoStream

Предлагаемый нами подход базируется на следующих предположениях, относящихся к публикации информации о новых событиях:

а) *минимальное время, прошедшее с момента* публикации (это предположение базируется на последовательном рассмотрении входящих в систему новостных документов, а также на анализе дат, указанных в текстах самих документов);

б) минимизация веса термов, входящих в документ, по частотному словарю, сформированному на основании анализа массива документов в рамках окна наблюдения (условие, аналогичное максимизации параметра IDF в векторно-пространственной модели);

в) максимизация суммарного веса термов, входящих в документ, по плюс-словарю (содержащему важные для содержания новостей слова типа: "теракт", "конфликт", "сенсация" и т. п.);

г) ранг "авторитетности" источника (как правило, определяемый экспертами).

Если ввести обозначения:

n величина окна наблюдения потока новостей;

D_l текущий документ;

D_n последний документ из окна наблюдения;

D_i i-й документ;

$PlusDic$ плюс-словарь;

$sim(D_i, D_j)$ мера близости документа i документу j;

$sim(D_i, PlusDic)$ мера близости документа i плюс-словарю,

то второе и третье условия (предположения) можно записать следующим образом:

$$sim(D_l, PlusDic) > \alpha,$$

$$\sum_{j=2}^{n} sim(D_l, PlusDic) > \beta,$$

где α и β эмпирически определяемые параметры. При этом, если рассматривать только высокоранговые источники, то эти два условия на практике оказываются вполне достаточными для выявления новых событий.

Мера близости $\sin(D_i, D_j)$ может быть определена традиционно для векторно-пространственной модели.

Пусть $D_i = \{w_{ik}\} = \{w : w \| D_i\}$ документ, рассматриваемый как множество термов "Bag of Words", $D_i + D_j = \{w : w \| D_i \mid w \| D_j\}$ объединение термов из документов D_i и D_j вектор размерности N.

Определим вектор $E_i = \{e_{ik}\}$ размерности N, соответствующий документу D_i, следующим образом:

$e_{ik} = 1$, если $w_{ik} \| D_i$,

$e_{ik} = 0$, иначе.