УДК 004.5

А. Г. Додонов[1], Д. В. Ландэ[2]
[1]Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина
[2]Информационный центр «ЭЛВИСТИ»
ул. М. Кривоноса, 2а, 03037 Киев, Украина

Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга

Приведены подходы к решению проблемы выявления фактографической информации из неструктурированных текстовых потоков. Описаны технологические решения, позволяющие извлекать из полнотекстовых документов такие понятия как фирмы, фамилии, географические названия и т.п., а также выявлять силу их взаимосвязей на основе применения двух алгоритмов. Первый из этих алгоритмов основывается на учете совместного вхождения понятий в одни и те же документы, а второй на учете общего для рассматриваемых понятий контекста.

Ключевые слова: *контент-мониторинг, информационный поток, выявление понятий, глубинный анализ текстов, взаимосвязь понятий.*

В настоящее время информационное пространство Интернет развилось до уровня, требующего новых подходов к поиску и анализу информации.

При проведении информационно-аналитических исследований на основе обработки информационного потока, формируемого в Интернет [1], особо актуальной оказывается задача автоматического извлечения из текстов фактографической информации [2]. При этом ввиду значительных объемов и динамики информационных потоков контент-анализ осуществляется сегодня с использованием современных информационно-аналитических систем.

Контент-мониторинг

Очевидно, следует признать, что изначальные парадигмы поисковых систем и контент-анализа, сформированные десятилетия тому назад, уже не отвечают реальной ситуации. Один из подходов к решению задачи извлечения фактов из текстовых документов и выявления их взаимосвязей базируется на технологии контент-мониторинга, который можно рассматривать как непрерывный во времени

содержательный анализ информационных потоков с целью получения необходимых качественных и количественных информационных срезов.

Именно непрерывная аналитическая обработка сообщений является самой характерной чертой этого подхода, который позволяет извлекать факты из тестов, выявлять новые понятия, формировать разнообразные статистические отчеты. Названные задачи сегодня охватываются двумя основными технологиями — извлечением фактографической информации из текстов (Information Extraction [2]) и глубинным анализом текстов (Text Mining [1]).

Современный уровень контент-мониторинга охватывает также задачи выявления взаимосвязей понятий, извлекаемых из документов, группировки этих понятий, визуализации. В этом случае на помощь приходят методы кластерного анализа, позволяющие на основе выявления латентных признаков формировать компактные группы понятий, выявлять главные из них, визуализировать взаимосвязи.

Названные задачи сегодня частично решаются ведущими контент-провайдерами во всем мире. Так, в 2006 году компания «Яндекс» в рамках своего новостного сервиса предоставила доступ к справочной информации о людях, упоминаемых в СМИ путем автоматического извлечения фактов из текстов и группировки их в пресс-портреты.

В компании «Интегрум-Техно» разработана автоматически пополняемая база данных, содержащая информацию о людях и организациях, связанных отношением «занимать должность» [3]. Основной принцип, используемый при выделении фактов, состоит в следующем: в предложении выделяются лексические единицы, указывающие на то, что в данном месте может встретиться группа «должности» или «компании», затем вокруг этих слов с помощью грамматик строятся определенные именные группы, в которых вершинами являются найденные слова.

Система контент-мониторинга [4] обеспечивает автоматизированный сбор информации с Web-сайтов в режиме реального времени, ее структурирование, группировку по семантическим признакам, а также тематическое избирательное распределение и предоставление доступа к информационным базам данных в поисковых режимах. Перспективным направлением развития технологии InfoStream также является контент-мониторинг, средствами которого обеспечивается решение задач формирования цепочек основных тематических сюжетов, дайджестов, извлечение фактов (понятий) из текстов, построение таблиц взаимосвязей и гистограмм распределения понятий.

Подходы к выявлению фактографических данных из документов

Следует отметить, что подходы к извлечению различных типов понятий из текстов существенно разняться как по контексту их представления, так и по структурным признакам. Так, для выявления принадлежности документа к тематической рубрике могут использоваться специальным образом составленные запросы на информационно-поисковых языках, включающих логические и контекстные операторы, скобки и т.д. Выявление географических названий предполагает использования таблиц, в которых кроме шаблонов написания этих названий используются коды стран, названия регионов и населенных пунктов. В качестве од-

ного из примеров рассмотрим алгоритм выявления названий фирм в текстах документов (рис. 1).

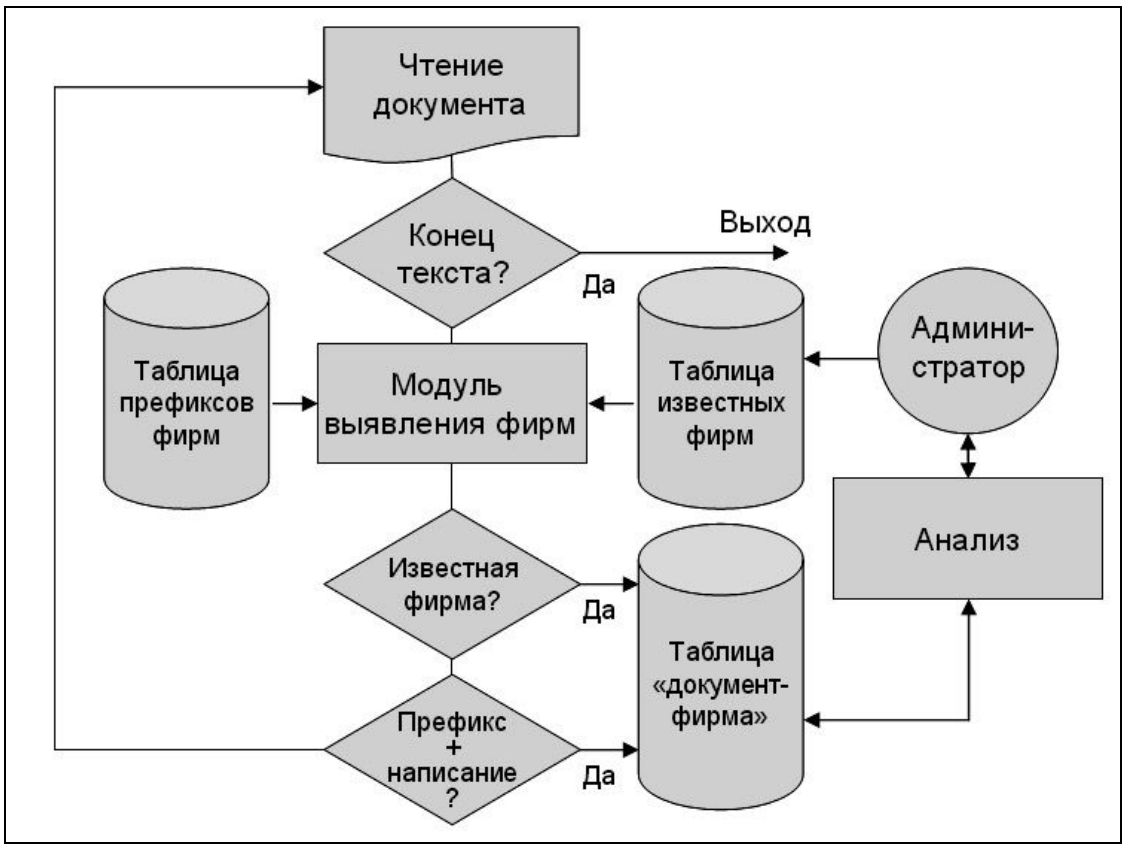

Рис. 1. Алгоритм выявления названий фирм из текстов документов

На вход системы поступает документ, который анализируется в процессе последовательного сканирования. Текст документа сравнивается с шаблонами, соответствующими названиям известных фирм, и если такие присутствуют, то они помещаются в специальную таблицу «документ–фирма». Также система извлечения фактографии предполагает выявление неизвестных изначально названий фирм на основании, как шаблонов, так и структурных исследований текста. При этом, в частности, используется таблица префиксов названий фирм, содержащая такие элементы, как «ООО», «ЗАО», «АО», «Компания» и др.

Выявленные понятия могут служить основой для построения многопрофильных информационных портретов или интерактивных ситуационных карт, соответствующих запросам пользователей [5]. Непосредственно по данным, представленным на ситуационной карте, отражающей наиболее актуальные понятия (термины, тематические рубрики, географические названия, имена персон, названия компаний) возможно выявление взаимосвязей понятий, т.е. сами ситуационные карты могут служить исходными данными для построения таблиц взаимосвязей.

Два подхода к построению таблиц взаимосвязей

Таблицы взаимосвязей понятий [6] строятся как статистические отчеты, отражающие близость (совместную встречаемость в новостных сообщениях или близость по сопутствующему контексту) отдельных понятий. Это симметричные матрицы, элементы которых — коэффициенты взаимосвязей понятий, соответствующих ее строкам и столбцам. Эти коэффициенты пропорциональны количеству документов входного информационного потока, которые одновременно соответствуют обоим понятиям, или количеству значимых лексических единиц, употребляемых совместно с данными понятиями. Таким образом, взаимосвязь понятий может быть оценена с помощью двух алгоритмов:

— совместного вхождения — путем расчета совместного вхождения этих понятий в одни и те же документы;

— контекстной близости — путем расчета корреляций наборов ключевых слов, входящих в документы, в которых упоминались данные понятия.

Рассмотрим формальное определение таблицы взаимосвязей понятий *TVP′*, построенной с помощью первого алгоритма. Обозначим p_j — понятие (j = 1,…, M); D_i — документ (i = 1,…, N); e_{ij} — признак соответствия понятия документу:

$$p_j \in D_i \Rightarrow e_{ji} = 1, \text{ иначе } e_{ji} = 0.$$

Можно определить уровень связи понятий p_j и p_k:

$$v'_{jk} = \sum_{i=1}^{N} e_{ji} e_{ki}.$$

Введя обозначение: $E = \| e_{ij} \|_{j=1,\ldots,M;\ i=1,\ldots,N}$, получаем:

$$TVP' = \| v'_{jk} \|_{j,\ k=1,\ldots,M}.$$

Для случая второго алгоритма, учитывающего контекстную близость, таблицу взаимосвязей понятий *TVP″* определим следующим образом. Обозначим p_j — понятие (j = 1,…, M); D_i — документ (i = 1,…, N); $\{w_1,\ldots, w_L\} = W_i$ — множество ключевых слов, входящих в D_i:

$$p_j \in D_i, \{w_1,\ldots,w_L\} = W_i \in D_i.$$

Введем понятие информационного портрета, как множества ключевых слов, соответствующих понятию p_j во всем массиве документов:

$$IP(p_j) = \bigcup_{\{i\ :\ p_j \in D_i\}} W_i.$$

Введем также понятие словаря системы

$$S = \| s_i \|_{i=1,\ldots,H}$$

и числовое множество $T(p_j)$ с элементами t_{ij}, соответствующее информационному портрету:

$$s_i \in IP(p_j) \Rightarrow t_{ji} = 1, \text{ иначе } t_{ij} = 0,$$

$$T(p_j) = \| t_{ij} \|_{i=1,\ldots,H}.$$

В этом случае уровень связи понятий p_j и p_k можно определить следующим образом:

$$v''_{jk} = (T(p_j), T(p_k)) = \sum_{i=1}^{H} t_{ij} t_{ik}.$$

Таким образом, таблица взаимосвязей понятий будет иметь вид:

$$TVP'' = \| v''_{jk} \|_{j,k=1,\ldots,M}.$$

Следует отметить, что таблица взаимосвязей первого вида всегда отражает взаимосвязи понятий точнее, чем таблица взаимосвязей второго типа, однако, таблица второго типа учитывает взаимосвязи более полно (рис. 2).

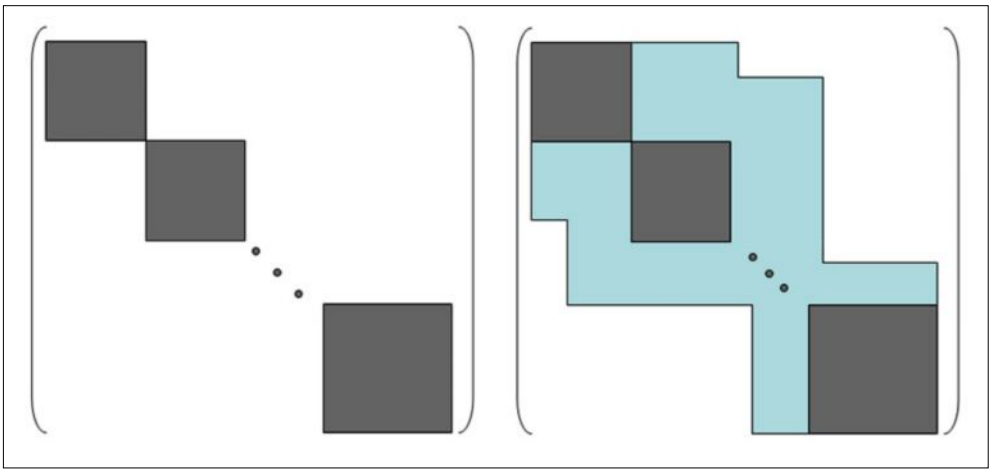

Рис. 2. Два варианта таблицы взаимосвязей понятий

Данное утверждение следует из теоремы, состоящей в том, что:

$$v'_{jk} > 0 \Rightarrow v''_{jk} > 0.$$

Действительно,

$$v'_{jk} > 0 \Rightarrow \exists i \ : \ p_j \in D_i, p_k \in D_i \Rightarrow$$

$$\bigcup_{\{w_l \in D_i\}} w_l \subset IP(p_j) \cap IP(p_k) \neq \emptyset \Rightarrow$$

$$(T(p_j),\ T(p_k)) = v''_{jk} > 0.$$

Утверждение, обратное данной теореме, в общем случае неверно. Проведем мысленный эксперимент, подтверждающий это замечание. Рассмотрим два понятия «пингвин» и «белый медведь». Эти понятия могут иметь ненулевое контекстное пересечение за счет таких ключевых слов, как «лед», «мороз», «рыба», однако понятие «пингвин» входит в документы, описывающие фауну Антарктики, а «белый медведь» — фауну Арктики.

Для переупорядочения понятий из таблицы взаимосвязей с целью выявления блоков — множеств наиболее взаимозависимых понятий (рис. 3) — применяются алгоритмы кластерного анализа. Покажем, как можно выделить некоторое число групп взаимосвязанных понятий методом *k*-means, который, как известно, является одним из самых эффективных для группировки динамических данных. Рассмотрим векторы-строки матрицы *TVP* – E_i (очевидно, ввиду симметричности матрицы *TVP* можно было бы рассматривать и столбцы). Простая задача оптимальной группировки векторов E_i в данном случае усложняется необходимостью при перестановке номеров векторов-строк одновременно переставлять соответствующие их компоненты для сохранения симметрии матрицы *E*.

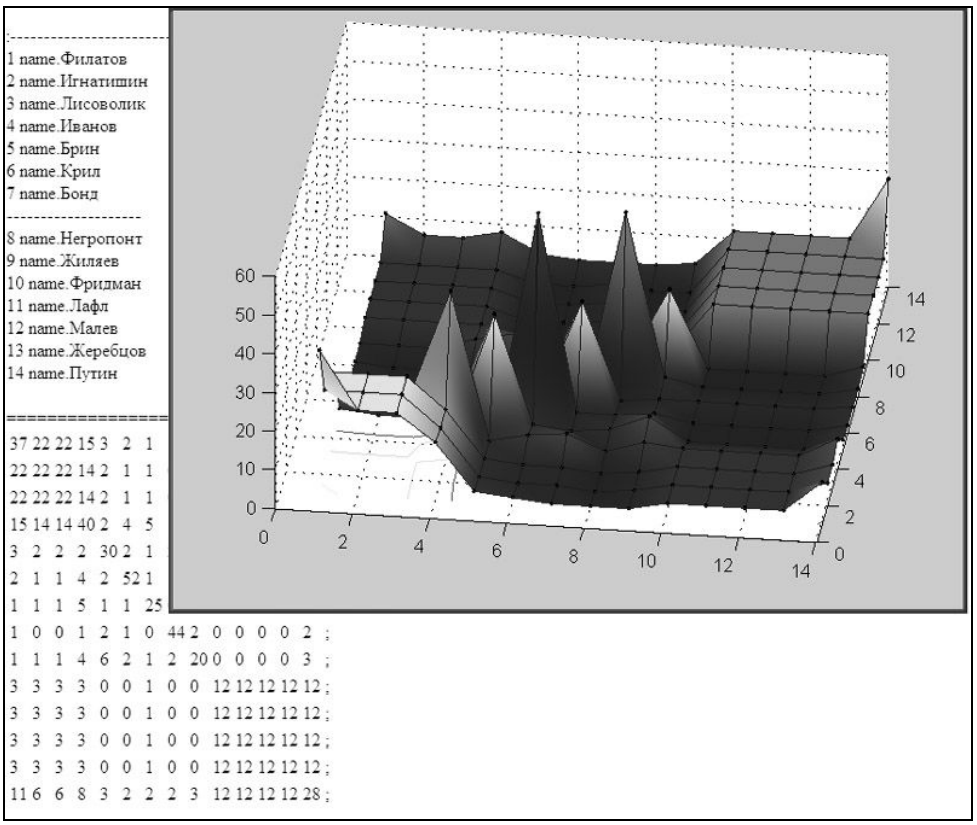

Рис. 3. Трехмерное представление взаимосвязи понятий

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 центроиду которого он наиболее близок. Затем строки-векторы группируются и перенумеровываются.

Для каждого из новых кластеров заново вычисляется 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 [7].

```
Произвольный выбор центроидов k-кластеров
while процесс формирования не стабилизировался do
        for каждого вектора-строки do
                найти центроид, наиболее близкий вектору-строке,
                приписать вектор-строку сответствующему  кластеру
        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 перенумерации
        end for
end while
```

Заключение

В качестве примеров современного применения технологии контент-мониторинга можно привести автоматическое выявление основных сюжетных цепочек, формирование рефератов и дайджестов, извлечение фактографий из текстов, выявление взаимосвязей понятий, автоматическую кластеризацию взаимосвязей для выявления наиболее важных из них.

Благодаря уже существующим возможностям систем контент-мониторинга, эта технология может способствовать значительному повышению качества информационно-аналитической работы. По сравнению с традиционными подходами использование технологии контент-мониторинга обеспечивает такие преимущества как получение оперативных количественных и качественных аналитических срезов по мере появления информации в Интернет, своевременное получение необходимой профильной фактографической информации при включении рабочих мест аналитиков в динамическое информационное пространство.

Вместе с тем, своего решения ждут проблемы автоматического выявления тональности взаимосвязей, в простейшем случае — определение принадлежностей взаимосвязей к положительным (группирующим) или отрицательным (антагонистическим). Также на данном этапе пока рассмотрены взаимосвязи лишь в рамках целостных документов, предполагается расширить анализ взаимосвязей понятий на отдельные их части.

1. *Ландэ Д.В.* Основы интеграции информационных потоков. — К.: Інжиніринг, 2006. — 240 с.

2. *Ralph Grishman.* Information extraction: Techniques and Challenges. In Information Extraction (International Summer School SCIE-97) // Springer-Verlag. — 1997.

3. *Гершензон Л. М., Ножов И. М., Панкратов Д. В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности. // Труды Международного семинара «Диалог'2005» (Звенигород, 1–6 июня 2005 г.). — М.: Наука, 2005.

4. *Додонов А.Г., Ландэ Д.В.* Организация сети информационных прокси-серверов // Реєстрація, зберігання і оброб. даних. — 2006. — Т. 8, № 3. — С. 24–31.

5. *Григорьев А.Н., Ландэ Д.В.* Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream. // Труды Международного семинара «Диалог'2005» (Звенигород. — 1–6 июня 2005 г.). — М.: Наука, 2005. — С. 109–111.

6. *Леліков Г.І., Сороко В.М., Григор'єв О.М., Ланде Д.В.* Моніторинг діяльності органів виконавчої влади із застосуванням комп'ютерної системи контент-аналізу електронних ЗМІ // Вісник державної служби України. — 2002. — № 2. — С. 72–78.

7. Ландэ Д.В. Некоторые методы анализа новостных информационных потоков // Научные труды Донецкого национального технического университета. Серия: Информатика, кибернетика и вычислительная техника (ИКВТ-2005). — Вып. 93. — Донецк: ДонНТУ, 2005. — С. 277–287.

Поступила в редакцию 07.11.2006