О ЦЕННОСТИ ИНФОРМАЦИИ

В. А. Ловицкий

Одной из наиболее сложных задач машинной обработки информации является автоматическое составление рефератов текстовых сообщений. Сложность формализации этого процесса обусловлена необходимостью выделять основное содержание текста, т. е. осуществлять его прагматический анализ. В настоящей статье рассматривается система, обладающая ограниченным объемом «знаний» и оценивающая входную информацию с точки зрения этих «знаний».

Ранее нами была предложена модель некоторых процессов переработки информации вербальной системой памяти и высказано предположение о структуре ее памяти [1]. Второй уровень памяти этой модели (память на последовательности слов) можно интерпретировать как систему, множество возможных состояний которой задано в виде конечного множества $\Xi = \{W_1, W_2, \ldots, W_q\}$. В каждый момент времени система характеризуется одним и только одним из этих состояний.

Пусть $H = \{F_1, F_2, \ldots F_h\}$ есть конечное множество последовательностей состояний, определяющее в процессе обучения пути переходов системы из одного состояния в другое и вероятности этих переходов. Таким образом, обучение системы сводится к анализу последовательностей состояний, заданных множеством H, и к учету результатов этого анализа. В процессе обучения формируется тот объем знаний, который позволяет системе осуществлять оценку входной информации.

Последовательности состояний $F_i (F_i \in H,\ i = 1, 2, \ldots, h)$ назовем **смысловыми.** Любая из этих последовательностей начинается с начального состояния. Задано конечное множество $\Psi (\Psi \subset \Xi)$ начальных состояний. Каждая последовательность заканчивается состоянием $W_k (W_k \in \Xi)$, называемым **конечным**, или **поглощающим**.

Функционирование системы заключается в последовательном переходе из одного состояния в другое, начиная с состояния W_{l_0} $(W_{l_0} \in \Psi)$. Каждый такой переход называется шагом процесса. Обозначим переход системы из состояния W_i в состояние W_j через $W_i \to W_j$. Начальное состояние задается посредством некоторого случайного выбора, дающего состояние W_{l_0} с вероятностью P_{l_0}. При переходе системы в состояние W_k ее функционирование заканчивается.

Рассмотрим функционирование системы при условии, что вероятность перехода системы из состояния W_i в состояние W_j зависит только от состояния W_i, из которого она исходит в процессе рассматриваемого перехода. Эта зависимость задается числами p_{ij}, представляющими вероятность того, что в данный момент времени система находится в состоянии W_j при условии, если предыдущий момент времени она находилась в состоянии W_i. Вероятности перехода p_{ij} могут быть расположены

в матрицу вероятностей перехода, у которой сумма элементов каждой строки равна единице, т. е.

$$\sum_{i=1}^{q}\sum_{j=1}^{q} p_{ij} = q,$$

где первый индекс означает номер строки, а второй — номер столбца.

Обозначим вероятность перехода системы из состояния W_{j_0} $(W_{j_0} \in \Psi)$ в состояние W_k за m шагов через $p_{j_0k}^{(m)}$. Вероятность того, что переход свершился за m шагов по некоторому фиксированному пути, равна

$$p_{j_0k}^{(m)} = p_{j_0} \prod_{i=0}^{m-1} p_{i_i i_{i+1}}, \qquad (1)$$

где

$$p_{i_{m-1}i_m} = p_{i_{m-1}k}.$$

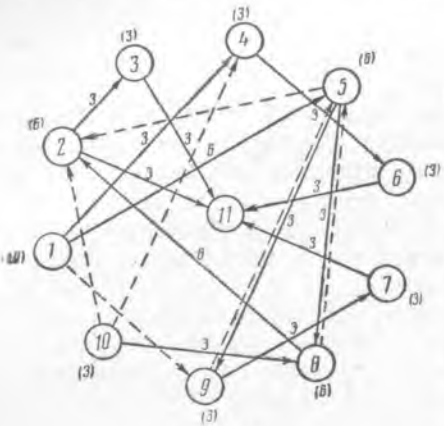

Рис. 1.

Будем считать, что состояние W_k достижимо из состояния W_{j_0}, если существует такое $m \geqslant 0$, что $p_{j_0k}^{(m)} > 0$.

Определение. *Множество состояний G называется замкнутым, если никакое состояние вне G не может быть достигнуто ни из какого состояния, входящего в G. Наименьшее замкнутое множество, содержащее G, называется замыканием G.*

Если некоторое состояние W_k образует замкнутое множество, то оно называется поглощающим состоянием, т. е. $p_{kk} = 1$. Мы будем рассматривать систему, имеющую только одно поглощающее состояние, равное W_k.

Таким образом, функционирование системы, при данном условии, можно рассматривать как марковский процесс с одним поглощающим состоянием.

Рассмотрим конкретную систему, множество возможных состояний которой задано в виде множества $\Xi = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$. Зададимся конечным множеством обучающих последовательностей $H = \{1 \to 5 \to 8 \to 2 \to 11,\ 1 \to 5 \to 9 \to 7 \to 11,\ 1 \to 4 \to 6 \to 11,\ 10 \to 8 \to 2 \to 3 \to 11\}$. Для данной системы множество начальных состояний имеет вид $\Psi = \{1, 10\}$, а конечное состояние равно 11. Каждому элементу множества H соответствует вес, определяющий число предъявлений системе при обучении данной последовательности. В табл. 1 указаны два распределения весов A и B для каждой последовательности. Систему, обученную заданным последовательностям с A распределением весов, назовем системой A, соответственно систему B распределением весов — системой B. На рис. 1 изображена система A, представленная в виде сети, в которой связи между элементами указаны сплошными линиями. Элементами этой сети служат возможные состояния данной системы.

В результате обучения элементы сети и возникшие между элементами связи приобрели определенный вес, который обозначим соответственно через $P_{эл}$ и $P_{св}$. Если от какого-то элемента сети отходит n связей, то для этого элемента $P_{эл} = \sum_{i=1}^{n} P_{св_i}$. Зная вес элемента W_i и вес связи между элементами W_i и W_j определим вероятность перехода

$W_i \to W_j$ следующим соотношением:

$$p_{ij} = \frac{P_{\text{св}\,i \to j}}{P_{\text{эл.}}}.$$

Матрицы вероятностей перехода систем A и B представлены в табл. 2.

Таблица 1

№ пп	Обучающие последовательности	Вес A	Вес B
1	$1 \to 5 \to 8 \to 2 \to 11$	3	5
2	$1 \to 5 \to 9 \to 7 \to 11$	3	3
3	$1 \to 4 \to 6 \to 11$	3	2
4	$10 \to 8 \to 2 \to 3 \to 11$	3	2

Вероятности послед. I и II	Система A	Система B
p_{oI}	$\frac{1}{6}$	$\frac{5}{14}$
p_{oII}	$\frac{1}{3}$	$\frac{3}{10}$

Таблица 2

Система A

Состояния / Состояния	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	$\frac{1}{3}$	$\frac{2}{3}$	0	0	0	0	0	0
2	0	0	$\frac{1}{2}$	0	0	0	0	0	0	0	$\frac{1}{2}$
3	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	1	0	0	0	0	0
5	0	0	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0
6	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	1
8	0	1	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	1	0	0	0	0
10	0	0	0	0	0	0	0	1	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1

Система B

Состояния / Состояния	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	$\frac{1}{5}$	$\frac{4}{5}$	0	0	0	0	0	0
2	0	0	$\frac{2}{7}$	0	0	0	0	0	0	0	$\frac{5}{7}$
3	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	1	0	0	0	0	0
5	0	0	0	0	0	0	0	$\frac{5}{8}$	$\frac{3}{8}$	0	0
6	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	1
8	0	1	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	1	0	0	0	0
10	0	0	0	0	0	0	0	1	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1

Будем считать, что объем «знаний» системы в какой-то определенной области задан в виде двух последовательностей: $1 \to 5 \to 8 \to 2 \to 11$ и

1→5→9→7→11). Эти последовательности обозначим соответственно через I и II.

На рис. 2 изображено дерево, позволяющее проследить все возможные пути переходов системы из начального состояния 1 в конечное состояние 11. Жирными линиями очерчены последовательности состояний I и II. Дерево построено по данным табл. 2.

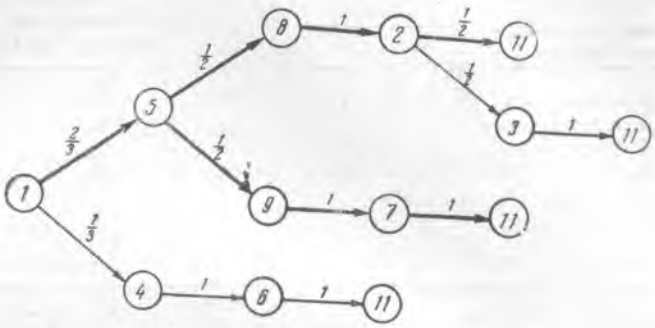

Рис. 2.

Задача обученной системы заключается в оценке предъявляемых последовательностей, т. е. в составлении «своего мнения». Меру ценности информации будем определять по Харкевичу [2] как двоичный логарифм отношения вероятностей достижения цели до и после получения информации, т. е.

$$\tilde{I} = \log_2 \frac{p_1}{p_0}, \qquad (2)$$

где $\tilde{I}$ — мера ценности входной информации;

p_1 — вероятность достижения цели после получения информации;

p_0 — вероятность достижения цели до получения информации.

Для нашей модели значения p_1 и p_0 будем определять по формуле (1), в которой число шагов m известно заранее и $p_0 = 1$, так как начальное состояние системы фиксировано.

Рассмотрим два варианта оценки системой входных последовательностей. При первом варианте как система A, так и система B поочередно оценивает каждую предъявляемую последовательность относительно последовательностей I и II, для которых значения p_{0I} и p_{0II} приведены в табл. 1. Предъявляемые последовательности будем считать смысловыми, т. е. принадлежащими некоторому конечному множеству смысловых последовательностей Θ, причем $H \subset \Theta$.

Системы A и B введены для того, чтобы показать зависимость возможностей системы от обучения. Суть обучения заключалась в том, что одним и тем же последовательностям уделялось различное внимание. Так, например, при обучении первой последовательности (см. табл. 1) система A повторила ее три раза, а система B — пять раз. Как сказались результаты обучения на возможностях систем, видно из табл. 3.

Первая предъявляемая последовательность (см. табл. 3) составлена из тех же элементов, что и последовательности I и II, но элементы в этой последовательности расположены в новом для I и II порядке. Каждая из последовательностей и система в целом оценивает эту последо-

вательность как отрицательную (см. табл. 3, значения $\tilde{I}_{\text{I}}$, $\tilde{I}_{\text{II}}$ и $\tilde{I}_{\Sigma}$). Величина I_{Σ} определяется как сумма $\tilde{I}_{\text{I}}$ и $\tilde{I}_{\text{II}}$, т. е. $\tilde{I}_{\Sigma} = \tilde{I}_{\text{I}} + \tilde{I}_{\text{II}}$.

Таблица 3

№ п/п	Предъявляемые последовательности	Система А					Система В				
		p_{I}	p_{II}	$\tilde{I}_{\text{I}}$	$\tilde{I}_{\text{II}}$	$\tilde{I}_{\Sigma}$	p_{I}	p_{II}	$\tilde{I}_{\text{I}}$	$\tilde{I}_{\text{II}}$	$\tilde{I}_{\Sigma}$
1	$1 \to 9 \to 5 \to 2 \to 11$	$\frac{36}{245}$	$\frac{27}{140}$	—0,20	—0,60	—0,80	$\frac{10}{33}$	$\frac{2}{11}$	—0,24	—0,72	—0,96
2	$10 \to 4 \to 6 \to 11$	$\frac{1}{6}$	$\frac{1}{3}$	0	0	0	$\frac{5}{14}$	$\frac{3}{10}$	0	0	0
3	$10 \to 8 \to 5 \to 9 \to 7 \to 11$	$\frac{6}{49}$	$\frac{8}{21}$	—0,45	+0,18	—0,27	$\frac{5}{16}$	$\frac{16}{45}$	—0,20	+0,26	+0,06
4	$10 \to 8 \to 2 \to 11$	$\frac{4}{21}$	$\frac{1}{3}$	+0,18	0	+0,18	$\frac{3}{8}$	$\frac{3}{10}$	+0,08	0	+0,08
5	$1 \to 5 \to 9 \to 7 \to 11$	$\frac{3}{20}$	$\frac{2}{5}$	—0,15	+0,26	+0,11	$\frac{25}{77}$	$\frac{4}{11}$	—0,15	+0,28	+0,13
6	$10 \to 2 \to 3 \to 11$	$\frac{1}{7}$	$\frac{1}{3}$	—0,22	0	—0,22	$\frac{5}{16}$	$\frac{3}{10}$	—0,20	0	—0,20

Последовательность будем называть положительной, отрицательной или нулевой, если мера ценности, определенная для этой последовательности, будет иметь соответственно знак плюс или минус либо будет равна 0. Условимся трактовать отрицательную меру ценности как несогласие системы с предъявленной последовательностью, а положительную — как согласие. Система может оценить положительно такую последовательность, которая, во-первых, представляет собой одну из последовательностей I или II (см. табл. 3 пятую последовательность), во-вторых, является новой последовательностью, т. е. включающей в себя как новые элементы, отличающиеся от элементов, из которых составлены последовательности I и II, так и новые связи между элементами (см. табл. 3 третью последовательность, предъявленную системе B).

Вторая последовательность, предъявляемая системе, включает в себя иные элементы, чем последовательности I и II. Эта последовательность оценивается системой как нулевая, т. е. не имеющая ничего общего с той областью «знаний», относительно которой и оцениваются все входные последовательности.

Третья последовательность наглядно показывает, как обучение системы определяет ее возможности. Так, система A «считает», что мера ценности данной входной последовательности отрицательна, а система B — что она положительна.

Четвертая, пятая, шестая и описанная выше третья последовательности приведены с целью показать, что «мнение» системы составляется в результате разрешения внутренних противоречий между последовательностями I и II.

Второй вариант оценки системой входных последовательностей заключается в том, что независимо от того, согласна или несогласна система с предъявленной последовательностью, она запоминает ее. Оценка последующих входных последовательностей осуществляется относительно тех же последовательностей I и II, для которых значения $p_{0\text{I}}$ $p_{0\text{II}}$ и $p_{1\text{I}}$, $p_{1\text{II}}$ меняются с запоминанием каждой новой входной последо-

вательности. Выражение (2) для определения, например, $\tilde{I}_1$ во втором варианте принимает следующий вид:

$$\tilde{I}_{1_{i+1}} = \log_2 \frac{p_{1_{i+1}}}{p_{1_i}},$$

где i — 0, 1, 2, 3, 4, 5.

На рис. 1 новые связи между элементами, возникшие в результате дообучения системы в процессе оценки входных последовательностей, изображены пунктирными линиями.

В табл. 4 приведены результаты второго варианта оценки входных последовательностей системами A и B. Обращает внимание тот факт, что система A, оценив первую последовательность как отрицательную, но затем запомнив ее, оценивает третью входную последовательность уже как положительную.

Таблица 4

№ п/п (i)	Предъявляемые последовательности	Система A					Система B				
		$p_{i\mathrm{I}}$	$p_{i\mathrm{II}}$	$\tilde{I}_{\mathrm{I}}$	$\tilde{I}_{\mathrm{II}}$	$\tilde{I}_{\Sigma}$	$p_{i\mathrm{I}}$	$p_{i\mathrm{II}}$	$\tilde{I}_{\mathrm{I}}$	$\tilde{I}_{\mathrm{II}}$	$\tilde{I}_{\Sigma}$
1	1 → 9 → 5 → → 2 → 11	$\frac{26}{245}$	$\frac{27}{140}$	−0,20	−0,60	−0,80	$\frac{10}{33}$	$\frac{2}{11}$	−0,24	−0,72	−0,96
2	10 → 4 → 6 → 11	$\frac{36}{245}$	$\frac{27}{140}$	0	0	0	$\frac{10}{33}$	$\frac{2}{11}$	0	0	0
3	10 → 8 → 5 → → 9 → 7 → 11	$\frac{27}{245}$	$\frac{6}{25}$	−0,42	+0,67	+0,25	$\frac{21}{88}$	$\frac{64}{275}$	−0,34	+0,37	+0,03
4	10 → 8 → 2 → 11	$\frac{63}{512}$	$\frac{6}{25}$	+0,15	0	+0,15	$\frac{224}{891}$	$\frac{64}{275}$	+0,07	0	+0,07
5	1 → 5 → 9 → → 7 → 11	$\frac{245}{2112}$	$\frac{175}{594}$	−0,10	+0,30	+0,20	$\frac{10}{297}$	$\frac{25}{88}$	−0,09	+0,28	+0,19
6	10 → 2 → 3 → 11	$\frac{245}{2376}$	$\frac{175}{594}$	−0,21	0	−0,21	$\frac{21}{99}$	$\frac{25}{88}$	−0,15	0	−0,15

В настояще время на основе использования изложенной выше методики расчета меры ценности информации составляется программа для машины «Урал-2», которая позволит существующей системе ЭВИВС (Элементарное Восприятие и Воспроизведение Слов) [1] оценивать отдельные входные предложения.

ЛИТЕРАТУРА

1. В. А. Ловицкий. Алгоритм ассоциативного поиска информации в вербальной системе памяти. II Украинская конференция по бионике. Тезисы докладов, К., 1968.
2. А. А. Харкевич. О ценности информации. Сб. «Проблемы кибернетики», вып. 4, М., 1960.