## *ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ*

*УДК 004.912*

*С. Ф. ЛИПНИЦКИЙ*

# АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВЫХ СООБЩЕНИЙ В ИНТЕРНЕТЕ НА ОСНОВЕ МОДЕЛИРОВАНИЯ ВЕРБАЛЬНЫХ АССОЦИАЦИЙ

*Объединенный институт проблем информатики НАН Беларуси*

*(Поступила в редакцию 31.09.2013)*

**Введение.** Актуальными направлениями исследований в области компьютерной лингвистики являются поиск тонально-окрашенной информации в Интернете и анализ ее тональности. Под тонально-окрашенными понимают текстовые сообщения, в которых выражено эмоциональное отношение их авторов к конкретным объектам или событиям. Своевременное выявление и аналитическая обработка такой информации позволяют минимизировать материальные потери или ущерб репутации, которые она может причинить персоне или организации, фигурирующей в тексте. Результаты анализа тонально-окрашенных интернет-сообщений полезны маркетологам, социологам и другим специалистам, изучающим общественное мнение с целью планирования своей деятельности.

Предложенный в данной статье подход к анализу тонально-окрашенной информации в отличие от существующих основан на использовании модели представления знаний на основе вербальных ассоциаций [1]. Эта модель позволяет учитывать наличие в текстах не только тонально-окрашенной лексики, но и семантических связей между словами, соответствующими ассоциативным отношениям между обозначаемыми ими сущностями предметной области. Благодаря этому модель обеспечивает формирование оценок тональности сообщений с использованием тонально-окрашенных корпусов текстов.

1. **Вербально-ассоциативные сети.** Вербально-ассоциативная сеть – это граф, вершинами которого являются слова, а ребрами – вербально-ассоциативные связи между ними [1]. Вершины сети помечены значениями информативности слов, а каждому ребру соответствует сила вербально-ассоциативной связи между словами. Такие сети могут быть построены для предложений, текстов и совокупностей текстов.

1.1. ***Тонально-окрашенные тематические корпусы текстов.*** Для оценки тональности сообщений в Интернете будем использовать совокупность тематических корпусов текстов, среди которых имеется корпус *Ct* с тонально-окрашенными текстами. Он состоит из совокупности тонально-окрашенных подкорпусов, каждому из которых соответствует некоторая оценка тональности. При *n*-бальной шкале оценок количество таких подкорпусов должно быть равно *n*, т. е. $Ct = \{Cp_i \mid i = \overline{1, n}\}$. Всякий *i*-й подкорпус $Cp_i$ включает в себя текстовые документы одинаковой тональности, т. е. корпус $Cp_i$ – пара $\langle Cp_i, Ev_i \rangle$, где $Ev_i$ – оценка тональности каждого документа из множества $Cp_i$.

1.2. ***Вербально-ассоциативное отношение.*** Обозначим через *W* множество всех слов некоторого языка *L*. Тогда отношение толерантности Θ (рефлексивное и симметричное бинарное отношение) на множестве *W* назовем *вербально-ассоциативным*, если любая упорядоченная пара слов (*a*, *b*) из множества *W* является элементом отношения Θ тогда и только тогда, когда слова *a*

и *b* из этой пары содержатся хотя бы в одном предложении языка *L*. Если пара (*a*, *b*) любых слов из множества *W* является элементом отношения Θ, т. е. (*a*, *b*) ∈ Θ, то (*a*, *b*) (по аналогии с [2]) будем называть *вербально-ассоциативной парой*.

1.3. ***Словари тонально-окрашенной лексики.*** Для поиска тонально-окрашенной информации и анализа ее тональности будем использовать следующие лингвистические словари.

*Частотный словарь тонально-окрашенной лексики* (табл. 1) $Dic_a = \{\langle a,\ n^a_{Cf}\ ,\ n^a_{\langle Cp_1,\ Ev_1\rangle}\ ,\ n^a_{\langle Cp_2,\ Ev_2\rangle}\ ,\ ...,\ n^a_{\langle Cp_n,\ Ev_n\rangle}\rangle \mid a \in W_{Cf}\}$. Здесь *a* – словоформа, $n^a_{Cf}$ и $n^a_{\langle Ct_i,\ Ev_i\rangle}$ ($i = \overline{1, n}$) – абсолютные частоты ее появления соответственно в полном корпусе текстов *Cf* и в *i*-м тонально-окрашенном подкорпусе, $W_{Cf}$ – множество всех словоформ полного корпуса. (Полный корпус текстов *Cf* – это объединение всех тематических корпусов.)

Таблица 1. **Фрагмент частотного словаря тонально-окрашенной лексики**

| Словоформа | $n^a_{Cf}$ | $n^a_{\langle Cp_1,\ Ev_1\rangle}$ | ... | $n^a_{\langle Cp_n,\ Ev_n\rangle}$ |
|---|---|---|---|---|
| ... | | | | |
| Качественный | 0204055 | 0056534 | ... | 0014445 |
| Качественного | 0401657 | 0074526 | ... | 0023747 |
| ... | | | | |

*Частотный словарь вербально-ассоциативных пар слов* (табл. 2), т. е. совокупность кортежей $Dic_{ab} = \{\langle (a, b),\ n^{ab}_{Cf}\ ,\ n^{ab}_{\langle Cp_1,\ Ev_1\rangle}\ ,\ n^{ab}_{\langle Cp_2,\ Ev_2\rangle}\ ,\ ...,\ n^{ab}_{\langle Cp_n,\ Ev_n\rangle}\rangle \mid a, b \in W_{Cf},\ n^{ab}_{Cf} \neq 0\ ,\ n^{ab}_{\langle Cp_i,\ Ev_i\rangle} \neq 0\ ,\ i = \overline{1, n}\}$, где $n^{ab}_{Cf}$ , $n^{ab}_{\langle Cp_i,\ Ev_i\rangle}$ – абсолютные частоты совместной встречаемости слов *a* и *b* в одном и том же предложении полного *Cf* и *i*-го тонально-окрашенного подкорпуса $Cp_i$ ($i = \overline{1, n}$).

Таблица 2. **Фрагмент частотного словаря вербально-ассоциативных пар слов**

| Пара словоформ | $n^{ab}_{Cf}$ | $n^{ab}_{\langle Cp_1,\ Ev_1\rangle}$ | ... | $n^{ab}_{\langle Cp_n,\ Ev_n\rangle}$ |
|---|---|---|---|---|
| ... | | | | |
| (Качественный, отсутствует) | 03020 | 00543 | ... | 00121 |
| (Предупреждение, избыточного) | 04023 | 00623 | ... | 00242 |
| ... | | | | |

*Словарь словоизменительных парадигм* [1] $Dic_{par} = \{(a, Par_a) \mid a \in W_{Cf},\ a \in Par_a\}$, где $Par_a$ – множество всех словоизменений слова *a* (включая *a*).

*Словарь синонимичных словоформ* [1] $Dic_{syn} = \{(a, Syn_a) \mid a \in W_{Cf},\ a \in Syn_a\}$, где $Syn_a$ – множество всех синонимов слова $a \in W_{Cf}$.

1.4. ***Задачи интернет-мониторинга тонально-окрашенной информации.*** В процессе интернет-мониторинга и анализа тонально-окрашенной информации решаются три основные задачи:

документальный поиск веб-страниц;

фактографический поиск на найденных страницах конкретной тонально-окрашенной информации, релевантной запросу;

анализ, оценка найденной тонально-окрашенной информации и составление отчета по результатам мониторинга.

Этим основным этапам интернет-мониторинга должно предшествовать построение вербально-ассоциативных сетей для тонально-окрашенных текстов; множеств тонально-окрашенных текстов; кратких сообщений; предложений; веб-страниц.

1.5. ***Вербально-ассоциативные сети тонально-окрашенных текстов.*** Пусть *T* – некоторый непустой текст языка *L*. Обозначим через $W_T$ множество всех слов текста *T*, а через $\Theta_T$ – сужение отношения Θ на множество $W_T$, т. е. $\Theta_T = \Theta \cap (W_T \times W_T)$. Обозначим через $G\Theta_T$ граф отношения $\Theta_T$. Пометим каждую вершину *a* графа $G\Theta_T$ значением информативности $I^a_T$ этого слова

(с учетом синонимии и словоизменения), а каждое ребро (*a*, *b*) – значением силы вербально-ассоциативной связи $I_T^{ab}$ слов *a* и *b* (также учитывая синонимию и словоизменения). Обозначим полученный граф через $Net_T$, который назовем *вербально-ассоциативной сетью* текста *T*. Приведем формулы для вычисления значений параметров $I_T^a$ и $I_T^{ab}$.

В [1] показано, что информативность слова в тексте вычисляется как отношение абсолютной частоты $m_T^a$ встречаемости слова в тексте *T* к абсолютной частоте $m_{Cf}^a$ его появления в полном корпусе текстов *Cf* (с учетом словоизменения и синонимии), т. е.

$$I_T^a = m_T^a / m_{Cf}^a .$$

Словоизменения зафиксированы в словаре словоизменительных парадигм $Dic_{par}$, а их синонимы – в словаре синонимичных словоформ $Dic_{syn}$. После нахождения синонимов каждого слова необходимо найти также их словоизменения в словаре $Dic_{par}$. С учетом этого обстоятельства общая формула для информативности $I_T^a$ примет вид

$$I_T^a = \frac{n_T^a + \sum\limits_{b\in Par_a,\, b\neq a} n_T^b + \sum\limits_{c\in Syn_a,\, c\neq a}\left(n_T^c + \sum\limits_{d\in Par_c,\, d\neq c} n_T^d\right)}{n_{Cf}^a + \sum\limits_{b\in Par_a,\, b\neq a} n_{Cf}^b + \sum\limits_{c\in Syn_a,\, c\neq a}\left(n_{Cf}^c + \sum\limits_{d\in Par_c,\, d\neq c} n_{Cf}^d\right)}, \qquad (1)$$

где $n_T^a$ и $n_{Cf}^a$ – абсолютные частоты встречаемости слова *a* в тексте *T* и полном корпусе текстов *Cf* (без учета словоизменения и синонимии); $n_T^b$ и $n_{Cf}^b$ – эти же частоты для словоформы *b*, входящей в состав парадигмы слова *a*; $n_T^c$ и $n_{Cf}^c$ – аналогичные параметры для синонима *c* слова *a*; $n_T^d$ и $n_{Cf}^d$ – абсолютные частоты встречаемости для словоформы *d*, входящей в состав парадигмы слова *c*.

Аналогичную формулу будем использовать для вычисления силы вербально-ассоциативной связи $P_T^{ab}$ слов *a* и *b*:

$$P_T^{ab} = \frac{n_T^{ab} + \sum\limits_{\substack{c\in Par_a,\, c\neq a\\ d\in Par_b,\, d\neq b}} n_T^{cd} + \sum\limits_{\substack{r\in Syn_a,\, r\neq a\\ s\in Syn_b,\, s\neq b}}\left(n_T^{rs} + \sum\limits_{\substack{p\in Par_r,\, p\neq r\\ q\in Par_s,\, q\neq s}} n_T^{pq}\right)}{n_{Cf}^{ab} + \sum\limits_{\substack{c\in Par_a,\, c\neq a\\ d\in Par_b,\, d\neq b}} n_{Cf}^{cd} + \sum\limits_{\substack{r\in Syn_a,\, r\neq a\\ s\in Syn_b,\, s\neq b}}\left(n_{Cf}^{rs} + \sum\limits_{\substack{p\in Par_r,\, p\neq r\\ q\in Par_s,\, q\neq s}} n_{Cf}^{pq}\right)}, \qquad (2)$$

где $n_T^{ab}$, $n_T^{cd}$, $n_T^{rs}$, $n_{Cf}^{ab}$, $n_{Cf}^{cd}$, $n_{Cf}^{rs}$ – абсолютные частоты совместной встречаемости пар слов (*a*, *b*), (*c*, *d*), (*r*, *s*) и (*p*, *q*) в одном и том же предложении текста *T* и полного корпуса текстов *Cf*.

При практической реализации системы интернет-мониторинга вербально-ассоциативную сеть $Net_T$ текста *T* представим в виде его поискового образа:

$$\text{ПО}_T = \{(a, I_T^a), (b, I_T^b), \ldots, ((c, d), P_T^{cd}), ((e, f), P_T^{ef}), \ldots \mid I_T^a, I_T^b, \ldots > I_T^0; P_T^{cd}, P_T^{ef}, \ldots > P_T^{00}\}, \qquad (3)$$

где *a*, *b*, *c*, *d*, *e*, *f*, ... – слова текста *T*; (*c*, *d*), (*e*, *f*), ... – вербально-ассоциативные пары слов; $I_T^a, I_T^b, \ldots$ – информативность слов *a*, *b*, ... текста *T*; $P_T^{cd}, P_T^{ef}, \ldots$ – сила вербально-ассоциативной связи между словами; $I_T^0$ и $P_T^{00}$ – пороговые значения информативности и силы связи.

1.6. ***Вербально-ассоциативные сети множеств тонально-окрашенных текстов.*** Пусть $\{T_i \mid i = \overline{1, n_T}\}$ – некоторое непустое множество тонально-окрашенных документов. Обозначим через *T* текст, полученный в результате объединения всех предложений каждого из текстов $T_i$. Тогда под вербально-ассоциативной сетью множества текстов $\{T_i \mid i = \overline{1, n_T}\}$ будем понимать сеть текста *T*. Параметры $I_T^a$ и $P_T^{ab}$ вычисляются по формулам (1) и (2), а поисковый образ текста *T* строится в соответствии с формулой (3).

1.7. ***Вербально-ассоциативные сети кратких сообщений и предложений.*** Вербально-ассоциативная сеть краткого сообщения *ST* – это подграф сети множества всех документов из тонально-окрашенного тематического корпуса текстов *Ct*. Вершины этого подграфа помечены значениями информативности соответствующих слов из корпуса текстов *Ct*, а ребра – показателями силы связи между словами из вербально-ассоциативной сети корпуса *Ct*.

Информативность слов краткого сообщения *ST* вычислим по формуле, аналогичной выражению (1):

$$I_{ST}^{a} = \frac{n_{Ct}^{a} + \sum\limits_{b \in Par_a,\, b \neq a} n_{Ct}^{b} + \sum\limits_{c \in Syn_a,\, c \neq a} \left( n_{Ct}^{c} + \sum\limits_{d \in Par_c,\, d \neq c} n_{Ct}^{d} \right)}{n_{Cf}^{a} + \sum\limits_{b \in Par_a,\, b \neq a} n_{Cf}^{b} + \sum\limits_{c \in Syn_a,\, c \neq a} \left( n_{Cf}^{c} + \sum\limits_{d \in Par_c,\, d \neq c} n_{Cf}^{d} \right)},$$

где $n_{Ct}^{a}$ и $n_{Cf}^{a}$ – абсолютные частоты встречаемости слова *a* из сообщения *ST* в корпусе текстов *Ct* и полном корпусе текстов *Cf* (без учета словоизменения и синонимии); $n_{Ct}^{b}$ и $n_{Cf}^{b}$ – эти же частоты для словоформы *b*, входящей в состав парадигмы слова *a*; $n_{Ct}^{c}$ и $n_{Cf}^{c}$ – аналогичные параметры для синонима *c* слова *a*; $n_{Ct}^{d}$ и $n_{Cf}^{d}$ – абсолютные частоты встречаемости для словоформы *d*, входящей в состав парадигмы слова *c*.

Для вычисления силы вербально-ассоциативной связи между словами краткого сообщения *ST* воспользуемся аналогом формулы (2):

$$P_{ST}^{ab} = \frac{n_{Ct}^{ab} + \sum\limits_{\substack{c \in Par_a,\, c \neq a \\ d \in Par_b,\, d \neq b}} n_{Ct}^{cd} + \sum\limits_{\substack{r \in Syn_a,\, r \neq a \\ s \in Syn_b,\, s \neq b}} \left( n_{Ct}^{rs} + \sum\limits_{\substack{p \in Par_r,\, p \neq r \\ q \in Par_s,\, q \neq s}} n_{Ct}^{pq} \right)}{n_{Cf}^{ab} + \sum\limits_{\substack{c \in Par_a,\, c \neq a \\ d \in Par_b,\, d \neq b}} n_{Cf}^{cd} + \sum\limits_{\substack{r \in Syn_a,\, r \neq a \\ s \in Syn_b,\, s \neq b}} \left( n_{Cf}^{rs} + \sum\limits_{\substack{p \in Par_r,\, p \neq r \\ q \in Par_s,\, q \neq s}} n_{Cf}^{pq} \right)},$$

где $n_{Ct}^{ab}$, $n_{Ct}^{cd}$, $n_{Ct}^{rs}$, $n_{Cf}^{ab}$, $n_{Cf}^{cd}$, $n_{Cf}^{rs}$ – абсолютные частоты совместной встречаемости пар слов (*a*, *b*), (*c*, *d*), (*r*, *s*) и (*p*, *q*) из сообщения *ST* в одном и том же предложении корпуса текстов *Ct* и полного корпуса *Cf*.

Поисковый образ краткого сообщения *ST* имеет вид

$$\text{ПО}_{ST} = \{(a, I_{Ct}^{a}), (b, I_{Ct}^{b}), ..., ((c, d), P_{Ct}^{cd}), ((e, f), P_{Ct}^{ef}), ... \mid I_{Ct}^{a}, I_{Ct}^{b}, ... > I_{Ct}^{0}; P_{Ct}^{cd}, P_{Ct}^{ef}, ... > P_{Ct}^{00}\},$$

где *a*, *b*, *c*, *d*, *e*, *f*, ... – слова сообщения *ST*; (*c*, *d*), (*e*, *f*), ... – вербально-ассоциативные пары слов из *ST*; $I_{ST}^{a}, I_{ST}^{b}, ...$ – информативность слов *a*, *b*, ... сообщения *ST*; $P_{T}^{cd}, P_{T}^{ef}, ...$ – сила вербально-ассоциативной связи между словами из *ST*; $I_{T}^{0}$ и $P_{T}^{00}$ – пороговые значения информативности и силы связи.

Предложение можно рассматривать как частный случай краткого сообщения. В связи с этим вербально-ассоциативная сеть и поисковый образ для него строятся так же, как и для сообщения *ST*.

1.8. ***Вербально-ассоциативные сети веб-страниц.*** Рассмотрим произвольную веб-страницу $S_{\text{веб}}$, вербально-ассоциативная сеть которой строится по аналогии с сетью тонально-окрашенного текста *T*. Значения информативности слов и вербально-ассоциативных пар в сети вычисляются по формулам (1) и (2) в предположении, что $T := S_{\text{веб}}$.

Поисковый образ страницы $S_{\text{веб}}$ аналогичен образу текста *T*:

$$\text{ПО}_{S_{\text{веб}}} = \{(a, I_{S_{\text{веб}}}^{a}); (b, I_{S_{\text{веб}}}^{b}), ..., ((c, d), I_{S_{\text{веб}}}^{cd}), ((e, f), I_{S_{\text{веб}}}^{ef}), ... \mid I_{S_{\text{веб}}}^{a}, I_{S_{\text{веб}}}^{b}, ... > I_{S_{\text{веб}}}^{0}; I_{S_{\text{веб}}}^{cd}, I_{S_{\text{веб}}}^{ef}, ... > I_{S_{\text{веб}}}^{00}\}.$$

2. **Анализ тонально-окрашенной информации.** Ему предшествуют поиск релевантных запросу веб-страниц с выдачей их адресов, а также поиск конкретных данных на этих страницах с представлением их информативных фрагментов.

2.1. ***Структура запроса на мониторинг тонально-окрашенной информации.*** Запрос пользователя на мониторинг тонально-окрашенной информации включает, как правило, два поля – название объекта мониторинга и его характеристики (табл. 3).

Таблица 3. **Пример запроса на мониторинг тонально-окрашенной информации**

| Объект мониторинга | Характеристика |
|---|---|
| Компания X. | Информация о компании. Выпускаемая продукция. Отзывы о качестве продукции. Отзывы о руководстве компании. |

В позиции объекта мониторинга тонально-окрашенной информации может быть указано физическое лицо, организация, средство массовой информации, событие и т. п. В качестве характеристик объекта мониторинга размещается любая текстовая информация о нем.

2.2. ***Этапы интернет-мониторинга.*** В общем случае мониторинг тонально-окрашенной информации реализуется в пять этапов.

На первом этапе индексируется первоначальный запрос пользователя.

На втором этапе корректируется поисковое предписание, полученное в результате индексирования этого запроса. Процесс коррекции реализуется следующим образом. По первоначальному поисковому предписанию проводится поиск релевантных документов в полном корпусе текстов. Найденное множество документов (динамический корпус текстов) индексируется и строится его вербально-ассоциативная сеть на основе полученного поискового образа. Эта сеть предъявляется пользователю, который ее корректирует путем исключения некоторых вершин. С использованием откорректированной вербально-ассоциативной сети строится новое поисковое предписание.

На третьем этапе проводится поиск веб-страниц по откорректированному поисковому предписанию. Найденные страницы анализируются на обновляемость. Из каждой обновленной вебстраницы извлекается вся текстовая информация.

На четвертом этапе выделяются информативные предложения на каждой найденной вебстранице. Строится контекст для всех информативных предложений. Полученные субтексты – это результат мониторинга.

На пятом этапе проводится анализ тональности каждого построенного субтекста с определением оценки тональности. Для этого выявляются наиболее релевантные тематические корпусы текстов. Оценка тональности субтекста совпадает с оценкой тональности релевантного ему корпуса.

Рассмотрим каждый из этапов интернет-мониторинга.

2.3. ***Индексирование запроса.*** Пусть $Z_1 = \{Z_1^{(1)}, Z_1^{(2)}\}$ – первоначальный запрос пользователя системы интернет-мониторинга, где $Z_1^{(1)}$ – наименование объекта мониторинга, а $Z_1^{(2)}$ – описание его характеристик.

Индексирование запроса $Z_1$ сводится к перечислению всех словоформ, присутствующих в цепочках $Z_1^{(1)}$ и $Z_1^{(2)}$, а также их синонимов и словоизменений. Полученное поисковое предписание представим в виде

$$\text{ПП}_1 = \{\text{ПП}_1^{(1)}, \text{ПП}_1^{(2)}\},$$

где $\text{ПП}_1^{(1)} = \{a \mid a \in Z_1 \cup Par_a \cup Syn_a\}$, $\text{ПП}_1^{(2)} = \{b \mid b \in Z_2 \cup Par_b \cup Syn_b\}$.

2.4. ***Коррекция поискового предписания.*** Исключим значения информативности слов и вербально-ассоциативные пары слов из поисковых образов всех документов полного корпуса текстов, т. е. преобразуем выражение (3) к виду

$$\text{ПО}_T = \{a, b, ... \mid a, b, ... \in T\}.$$

Сформируем динамический корпус текстов, релевантных запросу $Z_1$. С этой целью проведем двухшаговый поиск текстовых документов: сначала по запросу $Z_1^{(1)}$ в полном корпусе текстов *Cf*, а затем по запросу $Z_1^{(2)}$ в множестве текстов, найденных на первом шаге по запросу $Z_1^{(1)}$. В качестве критерия выдачи будем использовать мощность пересечения поисковых образов документов и поисковых предписаний.

Определим для каждого текста $T \in Cf$ мощность пересечения множеств $\text{ПО}_T$ и $\text{ПП}_1^{(1)}$, т. е. $m_1 = |\text{ПО}_T \cap \text{ПП}_1^{(1)}|$. Обозначим через $DZ_1^{(1)}$ множество документов полного корпуса текстов, найденных по поисковому предписанию $\text{ПП}_1^{(1)}$:

$$DZ_1 = \{T \mid T \in Cf, m_1 \geq m_1^0\},$$

где $m_1^0$ – пороговое значение мощности множества $DZ_1$.

В множестве документов $DZ_1$ проведем поиск документов по поисковому предписанию $\text{ПП}_1^{(2)}$. Обозначим результат поиска (динамический корпус текстов) через $DZ_2$:

$$DZ_2 = \{T \mid T \in DZ_1, m_2 \geq m_2^0\},$$

где $m_2 = |\text{ПО}_T \cap \text{ПП}_1^{(2)}|$.

Построим вербально-ассоциативную сеть корпуса $DZ_2$ как множества тонально-окрашенных текстов. Эту сеть предъявим пользователю для коррекции, т. е. для исключения из сети некоторых вершин, не отражающих содержание запроса. Использовав откорректированную сеть, создадим поисковый образ динамического корпуса текстов, который приобретает статус откорректированного поискового предписания:

$$\text{ПП}_2 = \{(a, I^a_{DZ_2}), (b, I^b_{DZ_2}), ..., ((c,d), P^{cd}_{DZ_2}), ((e,f), P^{ef}_{DZ_2}), ... \mid I^a_{DZ_2}, I^b_{DZ_2}, ... > I^0_{DZ_2}; P^{cd}_{DZ_2}, P^{ef}_{DZ_2}, ... > P^{00}_{DZ_2}\},$$

где $a$, $b$, $c$, $d$, $e$, $f$, ... – слова из поискового предписания $\text{ПП}_2$; $(c, d)$, $(e, f)$, ... – вербально-ассоциативные пары слов из $\text{ПП}_2$; $I^a_{DZ_2}, I^b_{DZ_2}, ...$ – информативность слов $a$, $b$, ... из $\text{ПП}_2$ в динамическом корпусе текстов $DZ_2$; $P^{cd}_{DZ_2}, P^{ef}_{DZ_2}, ...$ – сила вербально-ассоциативной связи между словами из $\text{ПП}_2$ в корпусе $DZ_2$; $I^0_{DZ_2}$ и $P^{00}_{DZ_2}$ – пороговые значения информативности и силы связи.

2.5. ***Поиск веб-страниц.*** Обозначим через $\text{ПП}'_2 = \{a, b, ..., (c, d), (e, f), ...\}$ поисковое предписание, полученное из предписания $\text{ПП}_2$ путем исключения из него показателей информативности слов и силы вербально-ассоциативной связи между ними. Точно так же преобразуем и поисковые образы всех веб-страниц.

Поиск веб-страниц реализуем в два этапа. На первом этапе проведем поиск по предписанию $\text{ПП}'_2$. При поиске будем использовать в качестве критерия выдачи мощность пересечения множеств $m = |\text{ПО}'_s \cap \text{ПП}'_2|$, где $\text{ПО}'_s$ – преобразованный поисковый образ веб-страницы $s \in S_{\text{веб}}$ ($S_{\text{веб}}$ – множество всех веб-страниц). Тогда в качестве результата поиска по предписанию $\text{ПП}'_2$ имеем множество веб-страниц

$$S_{\text{ПП}'_2} = \{s \mid s \in S_{\text{веб}}, m \geq m^0\},$$

где $m^0$ – пороговое значение мощности множества $S_{\text{ПП}'_2}$.

На втором этапе результаты поиска ранжируем с использованием критерия выдачи, основанного на вычислении косинуса угла между векторами поискового предписания и поискового образа документа. Рассмотрим этот критерий применительно к ранжированию веб-страниц.

Обозначим через $W$ множество всех слов и вербально-ассоциативных пар из полного корпуса текстов *Cf.* Построим $n$-мерное евклидово пространство $E$. Для этого лексикографически упорядочим все слова и вербально-ассоциативные пары из множества $W$, т. е. сформируем кортеж $W = \langle a_1, a_2, ..., a_n \rangle$, где $a_i$ – слово или вербально-ассоциативная пара слов из полного корпуса текстов. Для каждой веб-страницы $s \in S_{\text{веб}}$ построим вектор ее поискового образа в пространстве $E$: $\mathbf{F}_s = (I^{a_1}_{S_{\text{веб}}}, I^{a_2}_{S_{\text{веб}}}, ..., I^{a_n}_{S_{\text{веб}}})$. Аналогично представим вектор поискового предписания $S_{\text{ПП}'_2}$: $\mathbf{F}_{\text{ПП}'_2} = (J^{a_1}_{S_{\text{веб}}}, J^{a_2}_{S_{\text{веб}}}, ..., J^{a_n}_{S_{\text{веб}}})$.

При поиске тонально-окрашенной текстовой информации нужно учитывать то обстоятельство, что в Интернете индексируются не сами документы, а веб-страницы, на которых они расположены. Веб-страница может содержать более одного текста, а запрос пользователя, как правило, ориентирован на поиск именно текстовых документов. В связи с этим в векторе $\mathbf{F}_s$ приравняем к нулю все координаты (значения информативности и силы вербально-ассоциативной связи), соответствующие всем нулевым компонентам вектора $\mathbf{F}_{\text{ПП}'_2}$. Обозначим полученный вектор через $\mathbf{F}'_s$.

Тогда для ранжирования результатов поиска по поисковому предписанию $\text{ПП}'_2$ воспользуемся критерием выдачи

$$\cos\varphi = \frac{\mathbf{F}'_s \mathbf{F}_{\text{ПП}'_2}}{|\mathbf{F}'_s|\,|\mathbf{F}_{\text{ПП}'_2}|} = \frac{\sum_{i=1}^{n} I^{a_i}_{S_{\text{веб}}} J^{a_i}_{S_{\text{веб}}}}{\sqrt{\sum_{i=1}^{n} (I^{a_i}_{S_{\text{веб}}})^2}\sqrt{\sum_{i=1}^{n} (J^{a_i}_{S_{\text{веб}}})^2}}.$$

2.6. ***Поиск тонально-окрашенной информации на веб-страницах.*** Фактографический поиск тонально-окрашенных сообщений на найденных веб-страницах сводится к выделению в текстах сообщений информативных фрагментов, релевантных поисковому предписанию $\text{ПП}'_2$. Процедура включает два этапа.

На первом этапе вычисляются информативность всех слов на веб-странице, которые являются элементами множества $\text{ПП}'_2$, а также сила вербально-ассоциативной связи для всех пар слов и информативность всех предложений.

На втором этапе выявляется контекстное окружение информативных предложений в виде информативных фрагментов. Совокупность этих фрагментов является результатом фактографического поиска. Рассмотрим более подробно этапы его реализации.

Информативность $I^a_{S_{\text{веб}}}$ каждого слова *a* на веб-странице $S_{\text{веб}}$ вычислим по формуле (1), считая, что $I^a_{S_{\text{веб}}} = 0$, если $a \in \text{ПП}'_2$. Силу вербально-ассоциативной связи $I^{cd}_{S_{\text{веб}}}$ для всех пар слов (*c*, *d*) вычислим по формуле (2). При этом предполагаем, что $T := S_{\text{веб}}$. Если хотя бы одно из слов *c*, *d* не является элементом множества $\text{ПП}'_2$, то полагаем $I^{cd}_{S_{\text{веб}}} = 0$.

Обозначим через $V^{\pi}_{S_{\text{веб}}}$ множество всех слов, а через $W^{\pi}_{S_{\text{веб}}}$ – множество всех вербально-ассоциативных пар слов в предложении π на веб-странице $S_{\text{веб}}$. Тогда для вычисления информативности $I^{\pi}_{S_{\text{веб}}}$ предложения используем формулу, аналогичную выражению (5) из [1]:

$$I^{\pi}_{S_{\text{веб}}} = \frac{\sum\limits_{a \in V^{\pi}_{S_{\text{веб}}},\, a \in \text{ПП}'_2} I^a_{S_{\text{веб}}} + \sum\limits_{(c,\,d) \in W^{\pi}_{S_{\text{веб}}};\, c,d \in \text{ПП}'_2} I^{cd}_{S_{\text{веб}}}}{\sqrt{\sum\limits_{a \in V^{\pi}_{S_{\text{веб}}},\, a \in \text{ПП}'_2} \left(I^a_{S_{\text{веб}}}\right)^2 + \sum\limits_{(c,\,d) \in W^{\pi}_{S_{\text{веб}}};\, c,d \in \text{ПП}'_2} \left(I^{cd}_{S_{\text{веб}}}\right)^2}}.$$

При формировании контекстного окружения информативных предложений для вычисления силы вербально-ассоциативной связи между предложениями построим формулу на основе выражения для информативности $I^{\pi\rho}_{Q}$ из [1]:

$$I^{\pi\rho}_{S_{\text{веб}}} = \frac{\sum\limits_{c \in \pi,\, d \in \rho} I^{cd}_{S_{\text{веб}}}}{\sqrt{\sum\limits_{c \in \pi,\, d \in \rho} \left(I^{cd}_{S_{\text{веб}}}\right)^2}},$$

где π и ρ – произвольные предложения на веб-странице $S_{\text{веб}}$.

Информативность сообщения *Q*, полученного в результате построения контекста информативного предложения, будем вычислять по формуле, аналогичной выражению $I^{Sub}_{Q}$ [3]:

$$I^{Q}_{S_{\text{веб}}} = \frac{\sum\limits_{\pi \in Q} I^{\pi}_{S_{\text{веб}}}}{\sqrt{\sum\limits_{\pi \in Q} \left(I^{\pi}_{S_{\text{веб}}}\right)^2}}.$$

Если в тексте найдено несколько информативных предложений, то при выявлении их контекста каждое из неинформативных предложений ν, расположенное между парой информативных, будем относить к контекстному окружению такого предложения μ из этой пары, для которого значение силы связи $I^{\mu\nu}_{S_{\text{веб}}}$ наибольшее.

2.7. ***Оценка тональности интернет-сообщений.*** Пусть по-прежнему *Ct* – тематический корпус текстов с тонально-окрашенной лексикой, состоящий из *n* (по числу оценок в шкале оценивания) подкорпусов, т. е. $Ct = \{Cp_i \mid i = \overline{1, n}\}$. Каждый подкорпус $Cp_i$ состоит из текстов одинаковой тональности и представляет собой пару $\langle Cp_i, Ev_i \rangle$ ($Ev_i$ – оценка тональности для всех текстов из множества $Cp_i$). Пусть также *T* – текстовое сообщение, полученное в результате формирования контекстного окружения к некоторому найденному на веб-странице $S_{\text{веб}}$ информативному предложению. Построим вектор $\mathbf{F}_T$ сообщения *T* и векторы $\mathbf{F}_{Cp_i}$ ($i = \overline{1, n}$) по аналогии с построением векторов $\mathbf{F}_S$ и $\mathbf{F}_{\text{ПП}'_2}$. Для каждой пары ($\mathbf{F}_T$, $\mathbf{F}_{Cp_i}$) вычислим косинус угла между этими векторами по формуле

$$\cos\left(\widehat{\mathbf{F}_T, \mathbf{F}_{Cp_i}}\right) = \frac{\mathbf{F}_T \mathbf{F}_{Cp_i}}{|\,\mathbf{F}_T\,|\,|\,\mathbf{F}_{Cp_i}\,|}, \; i = \overline{1, n}.$$

Тогда сообщению $T$ будет соответствовать оценка тональности $Ev_i$ при такой величине $i$, при которой $\cos(\widehat{\mathbf{F}_T, \mathbf{F}_{Cp_i}})$ принимает наибольшее значение.

3. **Формирование отчета по результатам интернет-мониторинга.** В отчете по результатам интернет-мониторинга и анализа тонально-окрашенной информации каждому сообщению $Q$ поставим в соответствие адрес веб-страницы, на которой оно расположено, а также его информативность $I^{Q}_{S_{\text{веб}}}$ и числовое значение оценки тональности $Ev_i$ в принятой шкале (табл. 4).

Таблица 4. **Пример отчета по результатам мониторинга и анализа тонально-окрашенной информации**

| Тонально-окрашенное сообщение | Адрес веб-страницы | Информативность сообщения (%) | Оценка тональности |
|---|---|---|---|
| Сильные дожди, идущие в Центральной Европе несколько дней, привели к наводнениям в ряде регионов Австрии, Чехии, Германии и Польши. Проводится эвакуация населения, затруднено транспортное сообщение. | http:// vsekommentarii.com/ news/2013/06/02/9108226.htm | 45 | 4 (из 10) |
| Недовольные оправдательным приговором 28 обвиняемым в трагедии в Порт-Саиде подожгли здание Египетского футбольного союза и клуб отдыха офицеров полиции. | http:// rss.novostimira.com/ n_2361553.html | 64 | 3 (из 10) |

**Заключение.** Предложенный в статье подход может применяться для анализа тонально-окрашенных сообщений как в Интернете, так и в локальных базах данных. Благодаря использованию тонально-окрашенных тематических корпусов текстов и вербальных ассоциаций достигается универсальность программного обеспечения системы, т. е. его независимость от используемых входных языков и предметной области. Имеется возможность оперативного изменения шкалы оценок тональности анализируемых сообщений.

## Литература


1. *Липницкий С. Ф.* // Информатика. 2011. № 4. С. 21–28.
2. *Мартинович Г. А.* Вербальные ассоциации и организация лексикона человека // [Электронный ресурс]. Филологические науки. 1989. № 3. С. 39–45. Режим доступа: http:// lit. lib. ru/m/martinowich_g_a/02assfilnauk. shtml. Дата доступа: 7.08.2013.
3. *Липницкий С. Ф., Мамчич А. А.* // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. 2011. № 1. С. 72–81.



*S. F. LIPNITSKY*


**A PITCH ANALYSIS OF TEXT MESSAGES IN INTERNET BASED ON MODELING OF VERBAL ASSOCIATIONS**

## Summary


The problems of Internet monitoring of tone-colored information were formulated. An approach to search and analysis of tone-colored messages on the basis of verbal associations modeling was proposed. The questions of construction of verbal-associative networks for sentences, text and web pages were considered.