ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ

УДК 004.912

С. Ф. ЛИПНИЦКИЙ

АЛГОРИТМЫ РУБРИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ И КРАТКИХ СООБЩЕНИЙ В СИСТЕМЕ ИНФОРМАЦИОННОГО ИНТЕРНЕТ-МОНИТОРИНГА

Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь,
e-mail: lipn@newman.bas-net.by

Предлагается подход к автоматической рубрикации текстовых документов и кратких сообщений, основанный на использовании тематических корпусов текстов. Разработанные алгоритмы могут быть использованы при рубрикации неструктурированных текстов на различных входных языках. Для каждого языка «вручную» должен быть построен рубрикатор, каждой позиции которого ставится в соответствие поисковый образ релевантного ей корпуса текстов.

Ключевые слова: информативность слов, информационный мониторинг, корпус текстов, рубрикация текстовых документов.

S. F. LIPNITSKY

ALGORITHMS OF CATEGORIZATION OF TEXT DOCUMENTS AND SUMMARIES IN A SYSTEM OF INFORMATIONAL INTERNET-MONITORING

The United Institute of Informatics Problems of the National Academy of Sciences of Belarus,
Minsk, Belarus, e-mail: lipn@newman. bas-net. by

An approach to automatic categorization of text documents and summaries is proposed. This approach is based on the use of thematic text corpora. The developed algorithms can be used to categorize unstructured texts on different input languages. A rubricator for each language must be built «by hand», where each item of which is associated with the search image relevant to her corpus.

Keywords: the information content of words, information monitoring, body text, headings text documents.

Введение. Информационный мониторинг – технология систематического сбора и обработки информации с целью использования ее при принятии решений в различных предметных областях. Необходимость автоматизации процессов мониторинга связана с большими интеллектуальными и финансовыми затратами предприятий и организаций при «ручном» выполнении этих работ. Мониторинг информации в Интернете характеризуется двумя наиболее существенными аспектами. Первый аспект связан с поиском релевантных запросу пользователя текстов среди большого количества неупорядоченных интернет-данных, а второй – с анализом и рубрикацией найденных текстовых документов и сообщений.

Различают монотематические и политематические тексты, а также краткие сообщения. Под монотематическим понимают текст, посвященный единой тематике (например, монография). Политематический текст (например, сборник статей) является «склейкой» (конкатенацией) нескольких монотематических текстов. Краткое сообщение – текст объемом порядка 10000 слов и меньше (например, статья или новостное сообщение). Рубрикация монотематического текста сводится к его индексированию и поиску релевантного тематического корпуса текстов. При ру-

брикации политематического текста релевантная позиция рубрикатора ищется для каждого его раздела. При рубрикации кратких сообщений каждое из них целиком соотносится с релевантной рубрикой.

В данной статье предлагается подход к автоматической рубрикации неструктурированных текстовых документов и кратких сообщений, основанный на использовании тематических корпусов текстов, накопленных в соответствии с позициями рубрикатора. Для рубрикации текстов предлагается использовать разработанную автором модель представления знаний о предметной области на основе вербальных ассоциаций [1]. В отличие от существующих методов (см., например, [2]), основанных преимущественно на составляемых «вручную» списках ключевых слов, предлагаемый подход обеспечивает автоматическое индексирование документов и сообщений.

1. Понятие рубрикатора текстов и кратких сообщений

При построении рубрикатора каждой его позиции соотнесем название рубрики и поисковый образ (ПО) релевантного ей тематического корпуса текстов. Формализуем понятие рубрикатора.

1.1. ***Упорядоченные тематические корпусы текстов.*** Рассмотрим множество $CT = \{Ct_i \mid i = \overline{1, n}\}$ всех тематических корпусов текстов и их объединение, т. е. полный корпус текстов $Cf = \bigcup_{i=1}^{n_i} Ct_i$. Определим на множестве CT отношение строгого порядка (антирефлексивное и транзитивное бинарное отношение) $\prec$. Обозначим через $\prec^r$ редукцию $\prec^r = \prec \setminus \prec^2$ строгого порядка $\prec$. Содержательно редукция $\prec^r$ соответствует отношению подчинения позиций рубрикатора (например, *раздел-подраздел*, *подраздел-пункт* и т. д.), каждая из которых представлена релевантным ей тематическим корпусом текстов. Позиции рубрикатора имеют, как правило, названия (например: *Методы обработки; Инструмент; Материалы* и т. д.).

1.2. ***Создание поисковых образов тематических корпусов текстов.*** Поисковый образ любого тематического корпуса текстов Ct_i – совокупность ключевых слов *a*, *b*, ..., каждому из которых приписано значение показателя его информативности $I^a_{Ct_i}$, $I^b_{Ct_i}$, ... :

$$\text{ПО}_{Ct_i} = \{(a, I^a_{Ct_i}), (b, I^b_{Ct_i}), ... \mid a \in Ct_i, b \in Ct_i, I^a_{Ct_i} > I^0_{Ct_i}, I^b_{Ct_i} > I^0_{Ct_i}\}, \quad (1)$$

где $I^0_{Ct_i}$ – пороговое значение информативности.

Будем рассматривать корпус текстов Ct_i как единый текст T_i, полученный в результате последовательного объединения всех предложений всех текстов из множества Ct_i. Тогда ПО_{Ct_i} корпуса Ct_i получим в результате индексирования текста T_i. Информативность каждой словоформы *a* из текста T_i при его индексировании будем вычислять как отношение абсолютной частоты встречаемости словоформы *a* в тематическом корпусе текстов Ct_i к абсолютной частоте ее появления в полном корпусе текстов *Cf* [2]:

$$I^a_{Ct_i} = n^a_{Ct_i} / n^a_{Cf}. \quad (2)$$

Информативность $I^a_{Ct_i}$ слова *a* вычисляется с учетом словоизменений и синонимии, которые зафиксированы в следующих лингвистических словарях [3]:

частотный словарь словоформ $Dic_a = \{\langle a, n^a_{Cf}, n^a_{Ct_1}, n^a_{Ct_2}, ..., n^a_{Ct_n}\rangle \mid a \in W_{Cf}\}$, в котором каждой словоформе приписаны частоты ее встречаемости n^a_{Cf}, $n^a_{Ct_1}$, $n^a_{Ct_2}$, ..., $n^a_{Ct_n}$ во всех корпусах текстов (W_{Cf} – множество всех словоформ полного корпуса текстов *Cf*);

частотный словарь слабоинформативных словоформ $We_a = \{\langle a, n^a_{Cf}, n^a_{Ct_1}, n^a_{Ct_2}, ..., n^a_{Ct_n}\rangle \mid a \in W_{Cf}, n^a_{Ct_i} / n^a_{Cf} \le I_0, i = \overline{1, n}\}$, где I_0 – пороговое значение информативности словоформы;

частотный словарь вербально-ассоциативных пар слов $Dic_{ab} = \{\langle (a, b), n^{ab}_{Cf}, n^{ab}_{Ct_1}, n^{ab}_{Ct_2}, ..., n^{ab}_{Ct_n}\rangle \mid a, b \in W_{Cf}, n^{ab}_{Cf} \ne 0, n^{ab}_{Ct_i} \ne 0, i = \overline{1, n}\}$, 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 тематического корпусов текстов Ct_i ($i = \overline{1, n}$);

частотный словарь слабоинформативных вербально-ассоциативных пар слов $We_{ab} = \{\langle (a, b), n^{ab}_{Cf}, n^{ab}_{Ct_1}, n^{ab}_{Ct_2}, ..., n^{ab}_{Ct_n} \rangle \mid a, b \in W_{Cf}, n^{ab}_{Cf} \neq 0, n^{ab}_{Ct_i} \neq 0, n^{ab}_{Ct_i} / n^{ab}_{Cf} \leq I_{00}, i = \overline{1, n} \}$, где I_{00} – пороговое значение информативности вербально-ассоциативной пары слов;

словарь словоизменительных парадигм $Dic_{par} = \{(a, Par_a) \mid a \in W_{Cf}, a \in Par_a\}$, состоящий из пар ⟨*словоформа*, *парадигма*⟩. В позиции парадигмы Par_a представлены все словоизменения данной словоформы *a*;

словарь синонимичных словоформ $Dic_{syn} = \{(a, Syn_a) \mid a \in W_{Cf}, a \in Syn_a\}$, включающий в себя пары ⟨*словоформа*, *синонимичные словоформы*⟩, в которых каждой словоформе *a* соответствует множество ее синонимов Syn_a.

С учетом словоизменения и синонимии формула (2) примет вид

$$I^a_{Ct_i} = \frac{n^a_{Ct_i} + \sum\limits_{b \in Par_a, b \neq a} n^b_{Ct_i} + \sum\limits_{c \in Syn_a, c \neq a} (n^c_{Ct_i} + \sum\limits_{d \in Par_c, d \neq c} n^d_{Ct_i})}{n^a_{Cf} + \sum\limits_{b \in Par_a, b \neq a} n^b_{Cf} + \sum\limits_{c \in Syn_a, c \neq a} (n^c_{Cf} + \sum\limits_{d \in Par_c, d \neq c} n^d_{Cf})}. \quad (3)$$

Тогда процесс построения ПО (1) тематического корпуса текстов Ct_i сводится к последовательному вычислению информативности $I^a_{Ct_i}$ каждого слова *a* из текста T_i и проверке неравенств $I^a_{Ct_i} > I^0_{Ct_i}$.

1.3. ***Определение рубрикатора.*** Пусть *G* – орграф отношения $\prec^r$. Пометим каждую вершину орграфа *G* парой, включающей в себя название соответствующей рубрики и поисковый образ релевантного ей тематического корпуса текстов. Обозначим полученный орграф через Net_{cat}, назовем его рубрикатором текстовых документов и кратких сообщений.

2. Рубрикация монотематических текстовых документов

Рубрикация монотематических текстов сводится к поиску релевантной вершины рубрикатора Net_{cat} с наибольшим значением критерия выдачи. Рубрикация реализуется в три этапа.

На первом этапе индексируется подлежащий рубрикации текстовый документ как запрос пользователя на поиск рубрики.

На втором этапе корректируется поисковое предписание, полученное в результате индексирования этого текста. При коррекции по первоначальному поисковому предписанию проводится поиск релевантных документов в полном корпусе текстов. Найденное множество документов индексируется и строится его поисковый образ, т. е. откорректированное поисковое предписание.

На третьем этапе проводится поиск релевантных рубрик по откорректированному поисковому предписанию. Из найденных рубрик выбирается та, которой соответствует наибольшее значение критерия выдачи. К найденной рубрике относится исходный текст.

2.1. ***Индексирование рубрицируемого монотематического текста.*** Рассмотрим произвольный монотематический текст T ($T \in Cf$). ПО текста T – совокупность информативных (ключевых) слов, каждому из которых соответствует значение его информативности. Этот образ является предписанием на поиск релевантной вершины рубрикатора:

$$\text{ПО}_T = \{(a, I^a_T), (b, I^b_T), ... \mid a \in T, b \in T, I^a_T > I^0_T, I^b_T > I^0_T\}. \quad (4)$$

В множестве ПО_T ключевым словам *a* и *b* каждой пары соответствуют величины их информативности I^a_T и I^b_T в тексте *T*, превосходящие пороговое значение I^0_T.

При индексировании текста *T* информативность I^a_T произвольного слова *a* текста *T* будем вычислять по формуле, аналогичной выражению (3):

$$I^a_T = \frac{n^a_T + \sum\limits_{b \in Par_a, b \neq a} n^b_T + \sum\limits_{c \in Syn_a, c \neq a} (n^c_T + \sum\limits_{d \in Par_c, d \neq c} n^d_T)}{n^a_{Cf} + \sum\limits_{b \in Par_a, b \neq a} n^b_{Cf} + \sum\limits_{c \in Syn_a, c \neq a} (n^c_{Cf} + \sum\limits_{d \in Par_c, d \neq c} n^d_{Cf})}. \quad (5)$$

2.2. ***Коррекция поискового образа рубрицируемого монотематического текста.*** При коррекции ПО (4) нужно построить динамический корпус текстов, релевантных тексту *T*, а затем проиндексировать его. Полученный индекс – откорректированный ПО рубрицируемого текста.

Рассмотрим *l*-мерное евклидово пространство *E*. Для его построения лексикографически упорядочим все слова полного корпуса текстов *Cf*, т. е. сформируем кортеж $W_{Cf} = \langle c_1, c_2, ..., c_l \rangle$. Для ПО_Q (4) каждого текста *Q* из полного корпуса текстов *Cf* построим вектор в пространстве *E*:

$$\mathbf{F}_{\text{ПО}_Q} = (J_{c_1},\ J_{c_2}, \dots, J_{c_l}),$$

где $J_{c_1},\ J_{c_2}, \dots, J_{c_l}$ – значения информативности ключевых слов c_1, c_2, ..., c_l соответственно. (Компонента вектора $J_{c_k} = 1$, если слово c_k присутствует в ПО_Q текста *Q*, и $J_{c_k} = 0$ в противном случае.) Аналогично представим вектор ПО_T рубрицируемого текста *T*:

$$\mathbf{F}_T = (I_{c_1},\ I_{c_2}, \dots, I_{c_l}).$$

Как показано в [1], в качестве критерия выдачи целесообразно использовать косинус угла между векторами $\mathbf{F}_T$ и $\mathbf{F}_{\text{ПО}_Q}$:

$$\cos\varphi = \frac{\mathbf{F}_T \mathbf{F}_{\text{ПО}_Q}}{|\mathbf{F}_T|\,|\mathbf{F}_{\text{ПО}_Q}|} = \frac{\sum_{k=1}^{l} I_{c_k} J_{c_k}}{\sqrt{\sum_{k=1}^{l} I_{c_k}^2}\sqrt{\sum_{k=1}^{l} J_{c_k}^2}}. \qquad (6)$$

Если эта мера превышает некоторый порог $\cos\varphi_0$, то текст *Q* будем считать элементом создаваемого динамического корпуса текстов *Dt*:

$$Dt = \{Q \mid Q \in Cf, \cos\varphi > \cos\varphi_0\}.$$

Индексирование динамического корпуса текстов. При индексировании корпуса текстов *Dt* информативность каждого его слова вычисляется по формуле, аналогичной выражению (3):

$$I_{Dt}^a = \frac{n_{Dt}^a + \sum\limits_{b \in Par_a,\, b \neq a} n_{Dt}^b + \sum\limits_{c \in Syn_a,\, c \neq a} \left(n_{Dt}^c + \sum\limits_{d \in Par_c,\, d \neq c} n_{Dt}^d\right)}{n_{Cf}^a + \sum\limits_{b \in Par_a,\, b \neq a} n_{Cf}^b + \sum\limits_{c \in Syn_a,\, c \neq a} \left(n_{Cf}^c + \sum\limits_{d \in Par_c,\, d \neq c} n_{Cf}^d\right)}. \qquad (7)$$

Полученный в результате индексирования ПО динамического корпуса текстов *Dt* – откорректированное поисковое предписание на поиск ПО тематического корпуса текстов, релевантного динамическому корпусу *Dt*:

$$\text{ПО}_{Dt} = \{(a, I_{Dt}^a), (b, I_{Dt}^b), \dots \mid a \in Dt, b \in Dt, I_{Dt}^a > I_{Dt}^0, I_{Dt}^b > I_{Dt}^0\}. \qquad (8)$$

Искомому ПО будет соответствовать позиция рубрикатора, релевантная рубрицируемому тексту.

2.3. ***Поиск релевантной позиции рубрикатора.*** Нужно найти в множестве $CT = \{Ct_i \mid i = \overline{1, n}\}$ тематический корпус текстов Ct_i, релевантный поисковому предписанию (8). При поиске будем использовать критерий выдачи, аналогичный критерию (6):

$$\cos\psi = \frac{\mathbf{F}_{Ct_i} \mathbf{F}_{\text{ПО}_{Dt}}}{|\mathbf{F}_{Ct_i}|\,|\mathbf{F}_{Dt}|}, \qquad (9)$$

где $\mathbf{F}_{Ct_i}$ – вектор ПО тематического корпуса текстов Ct_i, а $\mathbf{F}_{\text{ПО}_{Dt}}$ – вектор ПО динамического корпуса текстов D_t.

Результатом поиска является тематический корпус текстов Ct_i с наибольшим значением критерия выдачи cos ψ, причем cos ψ > cos ψ_0 (cos ψ_0 – пороговое значение критерия выдачи). Если же релевантный тематический корпус не найден, то необходимы включение в рубрикатор новой позиции и формирование соответствующего тематического корпуса текстов.

2.4. ***Алгоритм рубрикации монотематических текстовых документов.*** В соответствии с алгоритмом индексируется входной текстовый документ и ищется релевантная полученному его ПО позиция рубрикатора.

А л г о р и т м 1. На входе алгоритма – монотематический текст *T*, на выходе – вершина рубрикатора *G*, помеченная поисковым образом тематического корпуса текстов, релевантным тексту *T*, а также названием соответствующей рубрики.

1. Вычислить информативность I_T^a каждой словоформы *a* текста *T* по формуле (5).
2. Проиндексировать текст *T*, т. е. сформировать его ПО в виде выражения (4).
3. Сформировать динамический корпус текстов $Dt = \{Q \mid Q \in Cf, \cos\varphi > \cos\varphi_0\}$, используя критерий выдачи (6).
4. Проиндексировать корпус текстов D_t, т. е. сформировать его ПО в виде выражения (8).
5. Искать позицию (вершину) рубрикатора Net_{cat}, релевантную $ПО_{Dt}$.
6. Если позиция рубрикатора найдена, то конец, иначе перейти к п. 7.
7. Выдать сообщение «Необходимо сформировать новый тематический корпус текстов». Конец.

3. Рубрикация кратких сообщений

Процедура рубрикации кратких сообщений аналогична процессу рубрикации монотематических текстов. Отличие в алгоритме индексирования рубрицируемого сообщения в используемом критерии выдачи.

3.1. ***Индексирование рубрицируемого краткого сообщения.*** Рассмотрим произвольное краткое сообщения *Nov* (*Nov* ∈ *Cf*). Индексирование сообщения *Nov* сводится к перечислению всех словоформ, присутствующих в нем (кроме слабоинформативных), а также их синонимов и словоизменений. Полученный ПО сообщения *Nov* представим в следующем виде:

$$ПО_{Nov} = \{a \mid a \in Nov \cup Par_a \cup Syn_a, a \notin We_a\}. \quad (10)$$

Для поиска синонимов слова *a* используем словарь синонимичных словоформ Dic_{syn}, а все словоизменения этого слова ищем в словаре словоизменительных парадигм Dic_{par}. Слабоинформативные слова краткого сообщения *Nov* представлены в словаре We_a.

3.2. ***Коррекция поискового образа краткого сообщения.*** По аналогии с коррекцией ПО монотематического текста коррекцию краткого сообщения реализуем путем построения динамического корпуса текстов *DNov*, релевантных сообщению *Nov*. В результате индексирования этого корпуса получим откорректированный ПО рубрицируемого текста.

Информативность вербально-ассоциативной связи между текстами. Обозначим через W_{Nov} множество всех слов ПО рубрицируемого сообщения *Nov*, а через W_Q – множество всех слов ПО произвольного текста *Q* из полного корпуса текстов *Cf*. Тогда информативность вербально-ассоциативной связи между сообщением *Nov* и текстом *Q* будем вычислять по формуле

$$I^{NovQ} = \frac{\sum\limits_{a \in W_{Nov},\, b \in W_Q} I^{ab}}{\sqrt{\sum\limits_{a \in W_{Nov},\, b \in W_Q} (I^{ab})^2}}. \quad (11)$$

Информативность I^{ab} вербально-ассоциативной связи слов *a* и *b* вычисляется по формуле, аналогичной выражению (5):

$$I^{ab} = \frac{n^{ab}_{Nov\cup Q} + \sum\limits_{\substack{c\in Par_a,\, c\neq a\\ d\in Par_b,\, d\neq b}} n^{cd}_{Nov\cup Q} + \sum\limits_{\substack{r\in Syn_a,\, r\neq a\\ s\in Syn_b,\, s\neq b}} (n^{rs}_{Nov\cup Q} + \sum\limits_{\substack{p\in Par_r,\, p\neq r\\ q\in Par_s,\, q\neq s}} n^{pq}_{Nov\cup Q})}{n^{ab}_{Cf} + \sum\limits_{\substack{c\in Par_a,\, c\neq a\\ d\in Par_b,\, d\neq b}} n^{cd}_{Cf} + \sum\limits_{\substack{r\in Syn_a,\, r\neq a\\ s\in Syn_b,\, s\neq b}} (n^{rs}_{Cf} + \sum\limits_{\substack{p\in Par_r,\, p\neq r\\ q\in Par_s,\, q\neq s}} n^{pq}_{Cf})}, \qquad (12)$$

где $n^{ab}_{Nov\cup Q}$, $n^{cd}_{Nov\cup Q}$, $n^{rs}_{Nov\cup Q}$, $n^{pq}_{Nov\cup Q}$, n^{ab}_{Cf}, n^{cd}_{Cf}, n^{rs}_{Cf}, n^{pq}_{Cf} – абсолютные частоты совместной встречаемости пар слов (a, b), (c, d), (r, s) и (p, q) в одном и том же предложении объединения $Nov \cup Q$ сообщения Nov и текста Q, а также полного корпуса текстов Cf.

Информативность I^{NovQ} – это мера релевантности текста Q рубрицируемому краткому сообщению Nov, т. е. критерий выдачи. Если эта мера превышает некоторый порог I^{00}, то текст Q будем считать элементом создаваемого динамического корпуса текстов:

$$DNov = \{Q \mid Q \in Cf,\, I^{NovQ} > I^{00}\}.$$

Индексирование динамического корпуса текстов. При индексировании корпуса текстов Dt информативность каждого его слова вычисляется по формуле, аналогичной выражению (3):

$$I^{a}_{DNov} = \frac{n^{a}_{DNov} + \sum\limits_{b\in Par_a,\, b\neq a} n^{b}_{DNov} + \sum\limits_{c\in Syn_a,\, c\neq a} (n^{c}_{DNov} + \sum\limits_{d\in Par_c,\, d\neq c} n^{d}_{DNov})}{n^{a}_{Cf} + \sum\limits_{b\in Par_a,\, b\neq a} n^{b}_{Cf} + \sum\limits_{c\in Syn_a,\, c\neq a} (n^{c}_{Cf} + \sum\limits_{d\in Par_c,\, d\neq c} n^{d}_{Cf})}. \qquad (13)$$

Полученный в результате индексирования ПО динамического корпуса текстов $DNov$ – откорректированное поисковое предписание на поиск ПО тематического корпуса текстов, релевантного динамическому корпусу $DNov$:

$$\text{ПО}_{DNov} = \{(a, I^{a}_{DNov}), (b, I^{b}_{DNov}), \ldots \mid a \in DNov, b \in DNov, I^{a}_{DNov} > I^{0}_{DNov}, I^{b}_{DNov} > I^{0}_{DNov}\}. \qquad (14)$$

Искомому поисковому образу будет соответствовать позиция рубрикатора, релевантная рубрицируемому краткому сообщению.

3.3. ***Алгоритм рубрикации кратких сообщений.*** Согласно алгоритму, индексируется входное краткое сообщение, корректируется его ПО путем формирования динамического корпуса текстов и реализуется поиск релевантной позиции рубрикатора.

А л г о р и т м 2. На входе алгоритма – краткое сообщение Nov, на выходе – позиция рубрикатора G.

1. Проиндексировать краткое сообщение Nov, т. е. построить его ПО в виде выражения (10).

2. Вычислить информативность вербально-ассоциативной связи I^{ab} по формуле (12) между словами $a \in W_{Nov}$ и $b \in W_Q$ для всех возможных пар слов (a, b).

3. Вычислить информативность вербально-ассоциативной связи I^{NovQ} по формуле (11) между сообщением Nov и текстами $Q \in Cf$ и сформировать динамический корпус текстов $DNov$.

4. Проиндексировать корпус текстов $DNov$, т. е. сформировать его ПО в виде выражения (14), вычисляя информативность слов по формуле (13).

5. Искать вершину графа Net_{cat}, которой соответствует тематический корпус текстов, релевантный поисковому предписанию ПО_{DNov}.

6. Если позиция рубрикатора найдена, то конец, иначе перейти к п. 7.

7. Выдать сообщение «Необходимо сформировать новый тематический корпус текстов». Конец.

4. Рубрикация политематических текстовых документов

Признаком политематичности текста является наличие в нем монотематических разделов. Каждый раздел в зависимости от его объема можно рассматривать как монотематический текст или краткое сообщение. Поэтому при рубрикации разделов политематического текста будем использовать рассмотренные выше два алгоритма.

А л г о р и т м 3. На входе алгоритма – политематический текст $Q = \langle R_1, R_2, \ldots, R_k \rangle$, состоящий из разделов $R_1, R_2, \ldots, R_k$, на выходе – позиции рубрикатора G для каждого раздела текста.

1. $i := 1$.

2. Если R_i – монотематический текст, то перейти к п. 3, если же R_i – краткое сообщение, то перейти к п. 4.

3. $T := R_i$. Выполнить алгоритм 1. Перейти к п. 5.

4. $Nov := R_i$. Выполнить алгоритм 2.

5. Если $i = k$, то конец, иначе перейти к п. 6.

6. $i := i + 1$. Перейти к п. 2.

Заключение. Предложенные в статье алгоритмы могут быть использованы при рубрикации неструктурированных текстов на различных входных языках. Для каждого языка «вручную» должен быть построен рубрикатор, каждой позиции которого ставится в соответствие поисковый образ релевантного ей тематического корпуса текстов. Процессы индексирования текстовых документов и сообщений, а также их рубрикации осуществляются автоматически.

Список использованной литературы

1. *Липницкий, С. Ф.* Модель представления знаний в информационных системах на основе вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2011. – № 4. – С. 21–28.

2. *Хачумов, М. В.* Задача кластеризации текстовых документов / М. В. Хачумов // Информационные технологии и вычислительные системы. – 2010. – № 2. – С. 42–49.

3. *Липницкий, С. Ф.* Индексирование текстовой информации на основе моделирования вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2012. – № 3. – С. 94–102.

4. *Липницкий, С. Ф.* Моделирование информационного поиска на основе динамических корпусов текстов / С. Ф. Липницкий, А. А. Мамчич // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. – 2011. – № 1. – С. 72–81.

Поступила в редакцию 28.05.2015