ИСПОЛЬЗОВАНИЕ МЕХАНИЗМА ЛОГОГЕНА МОРТОНА ДЛЯ ТЕРМИНОЛОГИЧЕСКОГО АНАЛИЗА ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Ломонос Я.Г.
Донецкий национальный университет, г. Донецк
кафедра компьютерных технологий
E-mail: oakim@dongu.donetsk.ua

Abstract

Lomonos Y.G. Use of the Morton's logogen mechanism for the terminology analysis of electronic documents. *E-documents fuzzy processing is certain in work. The improvement Morton's logogen formal description is offered. The sequential text processing algorithm by logogen is described.*

Введение

В настоящее время количество и содержимое электронных источников быстро растет. К таким источникам можно отнести электронные библиотеки, личные каталоги информации, базы данных электронного документооборота и Интернет. Последний из перечисленных источников информации является наиболее обширным. Важно отметить, что в настоящий момент количество пользователей электронных источников информации также быстро растет. Это приводит к понижению квалификации рядового пользователя.

Целью пользователя при работе с электронными документами является поиск в документах фрагментов текста, по смыслу удовлетворяющих запрос пользователя, то есть содержащих определенные термины.

Этот этап обработки информации в настоящее время поддерживается с помощью текстовых редакторов, которые предоставляют ограниченные возможности для поиска фрагментов текста: сопоставление конечной последовательности символов запроса с текстом. Такой механизм сложен так как не позволяет задавать более гибкие запросы на поиск в естественном для пользователя виде. Кроме того, такой механизм малоэффективен, так как при поиске термина необходимо учитывать изменение словоформ в связи с изменением числа, рода и падежа, а так же возможность изменения порядка следования входящих в термин словоформ. Существующие на текущий момент текстовые редакторы и процессоры не имеют средств решения этой проблемы.

Поиск фрагментов текста, кроме перечисленных проблем, осложняется еще наличием ошибок в источниках информации. Существующие на сегодняшний день приложения, выполняющие проверку синтаксиса, в большинстве своем не осуществляют редактуру автоматически, а предоставляют инструментарий для человека-оператора, что связано с затратами временных и других ресурсов [1,2,3].

Таким образом, проблема реализации интеллектуального поиска по терминам фрагментов текста, содержащего ошибки является актуальной проблемой.

В основу работы положена гипотеза, что человек, изучая электронный документ, способен выделять термины независимо от того, в какой форме они представлены и независимо от допущенных ошибок. Значит, электронный текст содержит достаточное количество информации для адекватной обработки, если применить когнитивные модели обработки текста [4,5].

Данная работа дополняет [3,6,7,8,10] и посвящена формализации терминологической разметки электронного документа, базирующейся на математическом аппарате нечетких систем, основанной на похожести символов. В представленной работе описаны механизмы информационной технологии предварительной разметки электронного текста, а именно

структуры морфологического и терминологического анализа с использованием нечетких моделей представления текста, основанных на теории логогена Мортона [4,8].

Представление текста в форме нечеткой характеристики

Задача терминологической разметки текста решается с учетом того, что в обрабатываемом тексте возможны ошибки, которые появляются либо в следствии невнимательности оператора ПК, либо в следствии особенностей программ распознавания текстовой информации.

Задачу предварительной разметки электронного текста можно разделить на три фазы: символьный анализ, морфологический анализ и терминологический анализ.

Каждая фаза обработки текста заключается в построении нечеткой характеристики текста своего уровня. Итого можно выделить четыре уровня представления электронного текста: первичный текст, нечеткая характеристика текста уровня символов, нечеткая характеристика текста уровня морфов (нечеткие морфы) и нечеткая характеристика текста уровня терминов (нечеткие термины).

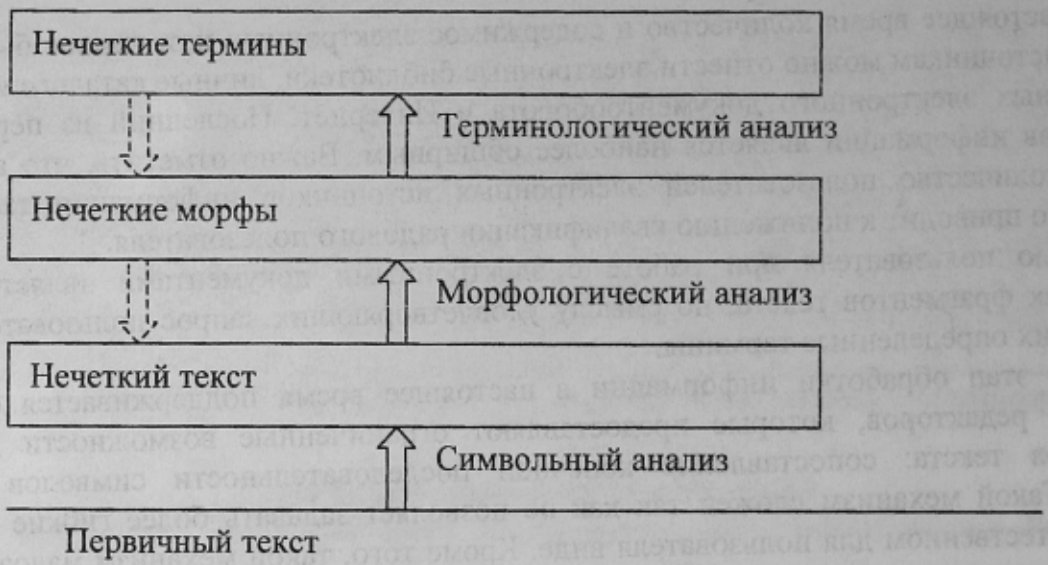

Рисунок 1 - Задача предварительной разметки нечеткого текста терминами

В данной работе предложено описывать все нечеткие величины, такие как похожесть двух символов алфавита, уверенность о наличии морфа и термина в тексте и уверенность о похожести двух символов алфавита друг на друга на базовом универсальном множестве фактора уверенности [9] $CF=[-1;+1]$, на котором задается некоторое подмножество $\underset{\sim}{S} = \{x \mid \mu_{\underset{\sim}{S}}(x), x \in CF\}$. Форма функции принадлежности нечеткого подмножества $\underset{\sim}{S}$ и её положение на оси CF служат нечёткой характеристикой гипотезы уверенности (либо в частном случае похожести).

Приближение центра тяжести данного множества к +1 усиливает доверие гипотезы (точечное подмножество $\underset{\sim}{S}^{+1} = \{+1 \mid 1\}$ описывает стопроцентную уверенность в истинности гипотезы), а к –1 – ее отрицание (точечное подмножество $\underset{\sim}{S}^{-1} = \{-1 \mid 1\}$ описывает стопроцентную уверенность в ложности гипотезы).

Для дефаззификации нечетких множеств в работе применяется метод центра тяжести, $p = \left\| \underset{\sim}{S} \right\|$, $p \in [-1;+1]$ – результат дефаззификации нечеткого множества.

Каждый уровень представления текста характеризуется своим алфавитом (словарем) символов. Алфавитом будем называть множество базовых символов каждого уровня: $S=\{s_1,s_2,...,s_n\}$. Так, в системе используются три алфавита: алфавит символов, алфавит морфов и алфавит терминов.

$S^0=\{s^0_1,s^0_2,...,s^0_n\}$ – алфавит символов украинского языка. В него входят все буквенные прописные и строчные символы украинского языка, цифры и символ «_», заменяющий все управляющие символы, которые могут встретиться в тексте, знаки препинания и символ пробела.

$S^1=\{s^1_1,s^1_2,...,s^1_r\}$ – алфавит (словарь) морфов украинского языка. Является объединением двух подсловарей морфов: словарь корней слов украинского словаря, и словарь вспомогательных морфов (префиксов и суффиксов), которых в украинском языке ограниченное количество.

$S^2=\{s^2_1,s^2_2,...,s^2_p\}$ – алфавит (словарь) терминов на украинском языке для некоторой области знаний.

Каждый символ алфавита S^1 и S^2 является последовательностью символов из алфавита предыдущего уровня:

$$s_i^j = s_{0_i}^{j-1} s_{1_i}^{j-1} \dots s_{u_i}^{j-1}. \quad (1)$$

В отличии от структуры морфа, который является условно произвольной последовательностью любых символов алфавита, термин имеет более строгую структуру, которую можно описать следующим видом:

$$s_i^2 = ([s_{1_i}^1]s_{2_i}^1[(s_{3_i}^1)]), \quad (2)$$

где $s_{1_i}^1$ – символ подсловаря морфов-префиксов, $s_{2_i}^1$ – символ подсловаря морфов-корней, $s_{3_i}^1$ – символ подсловаря морфов-суффиксов, выражение в квадратных скобках может отсутствовать, а выражение в круглых скобках повторяется не менее одного раза. В структуре термина участвуют только те морфы, которые остаются неизменными при склонении и изменении рода и числа. Основными морфами называются корневые морфы, а вспомогательными – префиксальные и суффиксальные. В данной работе принимается что все морфы, состоящие в структуре термина – корневые.

Уровни представления текста 2-4 являются нечеткими характеристиками текста и представляют собой конечную последовательность позиций, причем каждая позиция содержит исчерпывающую информацию о символах этого уровня, содержащихся в этой позиции. Эта информация представлена в нечетком виде на факторе уверенности CF [3,8,9].

Первичным текстом в данной работе называется конечная последовательность символов алфавита S^0:

$$A = s^0_{k_1} s^0_{k_2} \dots s^0_{k_T}, s^0_{k_i} \in S^0 \quad (3)$$

где $i=1...T$ – номер позиции символа в первичном тексте, а k_i – индекс символа из алфавита S^0, находящегося в i-той позиции.

Нечеткой характеристикой текста A уровней символов, морфов и терминов являются конечные последовательности позиций:

$$\begin{aligned}\widetilde{A} &= \widetilde{A}^1\widetilde{A}^2\ldots\widetilde{A}^T,\\ \widetilde{M} &= \widetilde{M}^1\widetilde{M}^2\ldots\widetilde{M}^T,\\ \widetilde{\Omega} &= \widetilde{\Omega}^1\widetilde{\Omega}^2\ldots\widetilde{\Omega}^T,\end{aligned} \qquad (4)$$

где T – количество позиций в тексте A;

каждая позиция нечетких характеристик $\widetilde{A}^i, \widetilde{M}^i, \widetilde{\Omega}^i$ – совокупность нечетких множеств $\{a^i_{\sim 1}, a^i_{\sim 2}, \ldots, a^i_{\sim n}\}$, характеризующих уверенность для каждого символа алфавитов S^0, S^1, S^2 в том, что в соответствующей позиции первичного текста был этот символ, морф, термин, $a^i_{\sim j} = \{x \mid \mu_{a^i_{\sim j}}, x \in CF\}$.

Фрагментом нечетких характеристик текста A (4) будем называть подпоследовательность позиций нечетких характеристик текста (4), начинающихся позицией t и заканчивающихся позицией $t+\tau$:

$$\begin{aligned}\widetilde{A}' &= \widetilde{A}^t\widetilde{A}^{t+1}\ldots\widetilde{A}^{t+\tau},\\ \widetilde{M}' &= \widetilde{M}^t\widetilde{M}^{t+1}\ldots\widetilde{M}^{t+\tau}.\end{aligned} \qquad (5)$$

Логоген Мортона

Для описания процесса обработки текста человеком Мортон разработал теоретическую модель описания слов. Формально логоген, как механизм интерпретации, представляет структуру с n входами и одним выходом. На вход логогена поступает последовательность ситуаций $\underset{\sim}{X} = \{\{X^1_{\sim 1}, X^1_{\sim 2}, \ldots, X^1_{\sim n}\}, \{X^2_{\sim 1}, X^2_{\sim 2}, \ldots, X^2_{\sim n}\}, \ldots, \{X^\tau_{\sim 1}, X^\tau_{\sim 2}, \ldots, X^\tau_{\sim n}\}\}$ – нечеткие символы, морфы, где $\underset{\sim i}{X^t}$ – фрагмент ситуации, представляющий t-тый вход логогена, $t=1,2,\ldots,\tau$. Выход логогена есть нечеткие уверенности о присутствии во входной последовательности морфа или термина. Задача логогена состоит в том, чтобы определить степень уверенности, с которой входная ситуация $\underset{\sim}{X}$ соответствует прототипу $y_k \in Y$, $k=1\ldots N$.

В теории логогена Мортона отмечается, что появление нового свидетельства при рассмотрении некоторой гипотезы изменяет уверенность, сформированную ранее на основе предыдущих свидетельств [4]. Это изменение является нелинейной функцией от значения уверенности вновь поступившего свидетельства и значения накопленной ранее уверенности гипотезы. В теории фактора уверенности [9] предлагается в качестве операции накопления уверенности использовать операцию

$$\Lambda(x_1, x_2) = \begin{cases} x_1 + x_2 - x_1 \cdot x_2, & x_1 \ge 0, x_2 \ge 0 \\ x_1 + x_2 + x_1 \cdot x_2, & x_1 < 0, x_2 < 0, \\ \dfrac{x_1 + x_2}{1 - MIN(x_1, x_2)} \end{cases} \qquad (6)$$

$$\Lambda^{\mu}(\mu_1, \mu_2) = \mu_1 + (\mu_2 - \mu_1) - \mu_1 \cdot (\mu_2 - \mu_1)$$

где x_1 – накопленная ранее уверенность в гипотезе, x_2 – значение уверенности вновь поступившего свидетельства, μ_1 – функция принадлежности накопленной ранее уверенность

в гипотезе, μ_2 – функция принадлежности уверенности вновь поступившего свидетельства. Аргументы x_1 и x_2 и результат операции $\Lambda(x_1,x_2)$ определены на факторе уверенности CF.

На рис. 2 проиллюстрирована работа операторов (6). Изображенные на этом рисунке нечеткое множество $\tilde{X}_1$ – накопленная ранее уверенность в гипотезе, $\tilde{X}_2$ – значение уверенности вновь поступившего свидетельства, а $\tilde{X}_3$ – результат выполнения операторов (6) над множествами $\tilde{X}_1$ и $\tilde{X}_2$.

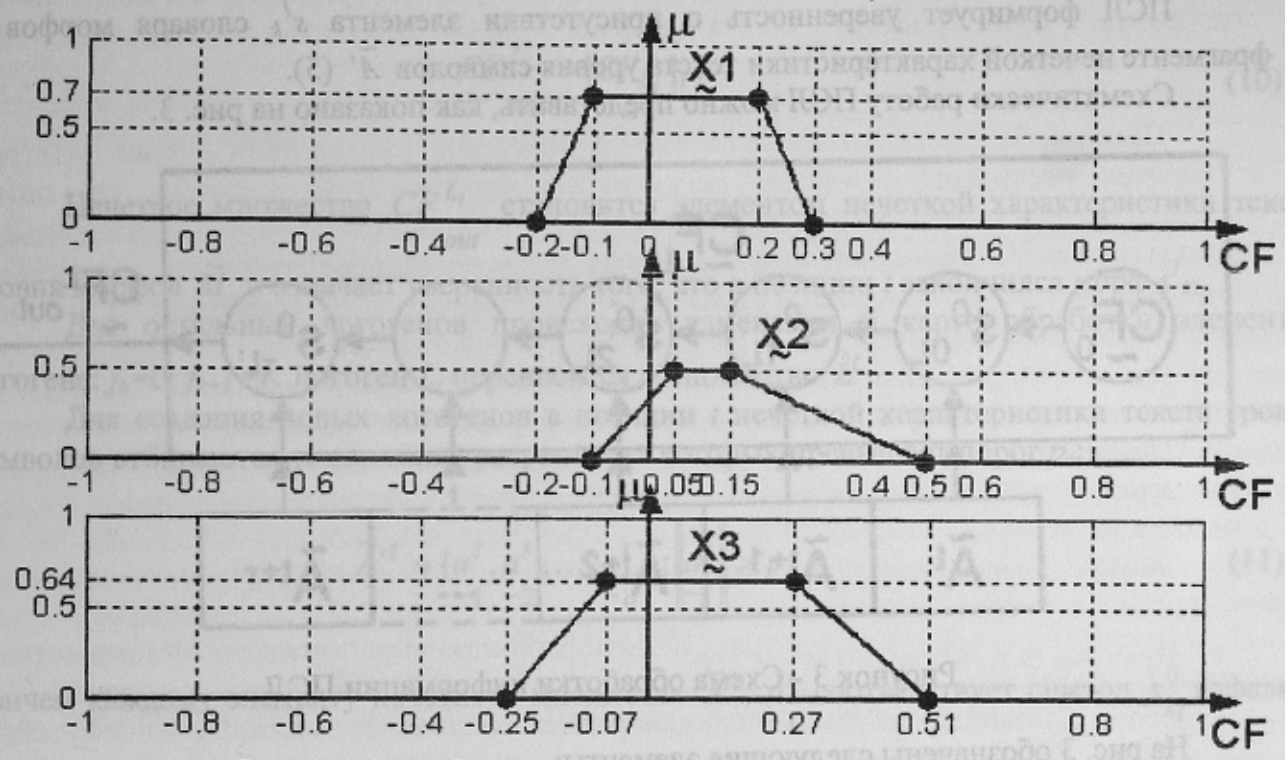

Рисунок 2 - Работа операторов накопления уверенности

В результате вывода на выходе логогена получаем уверенность в том, что входная ситуация описывается прототипом y_k, представляющая нечеткий морф или термин и представлена в виде нечеткого интервала на факторе уверенности.

Для дефаззификации полученных нечетких интервалов целесообразнее всего использовать метод центра тяжести нечеткого множества.

Для обработки текстовой информации автором была предложена усовершенствованная модель логогена и разработано два алгоритма работы логогена.

Логоген представляет собой структуру вида:

$$L = (s_k^j, map_L, \tilde{CF}_L), \qquad (7)$$

где s^j_k – обрабатываемый логогеном прототип элемента текста (морф или термин), являющийся последовательностью из u_k элементов словаря S^{j-1} (1); map_L – карта обработки элементов логогена: $map_L = f_1 f_2 \dots f_{u_k}$, f_i=1 если i-тый элемент последовательности (1) может быть обработан на следующем шаге обработки и f_i=0 в ином случае; $\tilde{CF}_L$ – внутреннее состояние логогена, характеризующий накопленную уверенность.

Обработка логогена может идти двумя способами:

1. последовательная n-тактная синхронная обработка логогена (ПСЛ) – применяется для формирования уверенности наличия морфемы из n символов в нечеткой характеристике текста уровня символов [10];
2. параллельная асинхронная обработка логогена из m элементов (ПАЛ) применяется для формирования уверенности наличия термина из m корневых морфов в нечеткой характеристике текста уровня морфов.

В рамках данной работы описывается работа логогена на примере последовательной n-тактной синхронной обработки.

Последовательная n-тактная синхронная обработка логогена на этапе морфологического анализа

ПСЛ формирует уверенность о присутствии элемента s^l_k словаря морфов S^l в фрагменте нечеткой характеристики текста уровня символов $\tilde{A}^t$ (5).

Схематически работу ПСЛ можно представить, как показано на рис. 3.

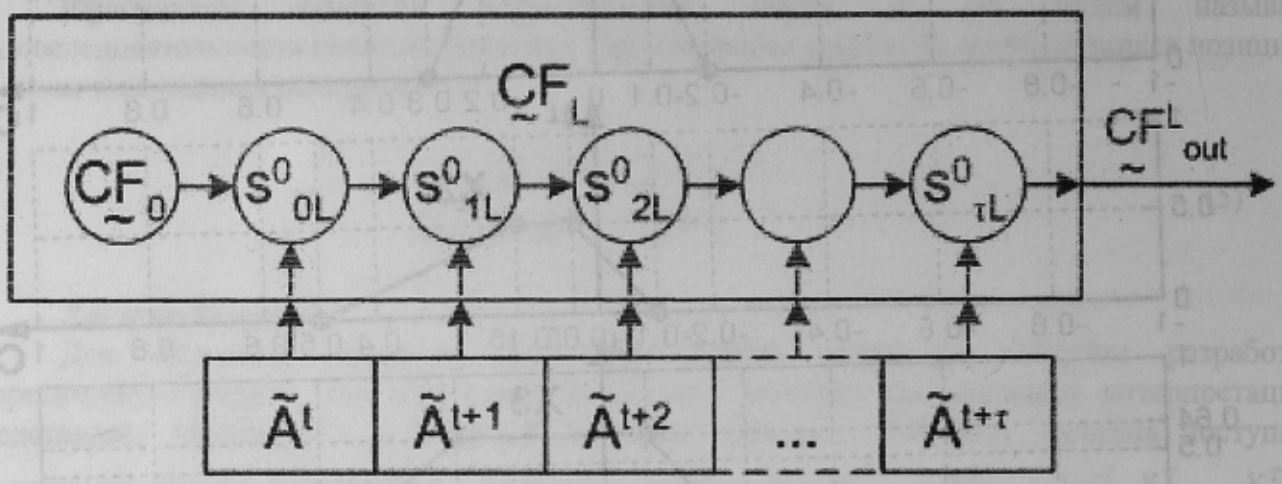

Рисунок 3 - Схема обработки информации ПСЛ

На рис. 3 обозначены следующие элементы:

$\tilde{A}^i$ – позиция фрагмента нечеткой характеристики текста уровня символов $\tilde{A}^t$ (5); CF_0 – начальное состояние логогена; s^0_{iL} – элемент структуры логогена, соответствующий структурному элементу обрабатываемого морфа s^l_{kL}; CF^L_{out} – выход логогена.

К началу обработки позиции *t* имеются следующие данные:

$\hat{L}^{t-1} = \{L_1, \ldots, L_{m_t}\}$ – множество логогенов, соответствующих обработанным в позиции *t-1* морфам.

Для каждого L_i происходит обработка символа $s^0_{k_i}$ последовательности (1) для которого f_k=1 и происходит накопление уверенности по формулам (6), добавляя к уже накопленной уверенности $[CF_L]_{t-1}$ уверенность символа $s^0_{k_i}$ в текущей позиции нечеткой характеристики текста уровня символов $a^t_{k_i} \in \tilde{A}^t$ с учетом важности этого символа – ω_{ki}:

$$[CF_{L_i}]_t = \Lambda([CF_{L_i}]_{t-1}, \omega_{k_i} \cdot a^t_{k_i}). \quad (8)$$

Операция $\omega \cdot \underset{\sim}{a}$ является масштабированием нечеткого множества $\underset{\sim}{a}$, по оси уверенности CF на величину ω:

$$\omega \cdot \underset{\sim}{a} = \{\omega \cdot x \,|\, \mu_{\underset{\sim}{a}}(x), x \in CF, \omega \in (0;1]\}. \quad (9)$$

Для тех логогенов, для которых $k = u$ (обрабатывается последняя позиция) в позиции t происходит окончательное формирование уверенности:

$$\underset{\sim}{CF}^{L_i}_{out} = [\underset{\sim}{CF}_{L_i}]_t, \quad (10)$$

Нечеткое множество $\underset{\sim}{CF}^{L_i}_{out}$ становится элементом нечеткой характеристики текста уровня морфов $\tilde{M}$ и означает уверенность того, что в позиции t закончился морф s^1_{kL}.

Для остальных логогенов происходит изменения в карте обработки элементов логогена: f_k=0; f_{k+1}=1. Логоген L_i переносится в множество $\hat{L}^t$.

Для создания новых логогенов в позиции t нечеткой характеристики текста уровня символов отбираются те символы, уверенность которых превышает порог p_s:

$$\tilde{A}'^t = \{\underset{\sim 1}{a^t}, \underset{\sim 2}{a^t}, \ldots, \underset{\sim q}{a^t}\} : \left\| \underset{\sim i}{a^t} \right\| > p_s, \quad (11)$$

причем каждому элементу нечеткого множества $\tilde{A}'^t$ $\underset{\sim i}{a^t}$ соответствует символ $s^0_{k_i}$ алфавита S^0.

Для каждого $s^0_{k_i}$ из словаря морфов S^1 отбираются те морфы, в структуре (1) которых присутствует символ $s^0_{k_i}$ на любой позиции, кроме последней:

$$S'^1 = \{s^1_j : s^1_j = s^0_{0_j} s^0_{1_j} \ldots s^0_{n_j-1} \wedge \exists v_j : s^0_{v_j} = s^0_{k_i}\}. \quad (12)$$

Если в множестве $\hat{L}^t$ нет такого логогена L_c, для которого $s^1_{k_i} = s^1_{k_c}$ и $f_{v_j+1} = 1$, создается логоген L_j, внутренне состояние которого вычисляется по формуле (6) между начальным состоянием ПСЛ и уверенностью символа $s^0_{k_i}$ с учетом важности этого символа:

$$[\underset{\sim}{CF}_{L_j}]_t = \Lambda(CF_0, \omega_{k_i} \cdot \underset{\sim k_i}{a^t}). \quad (13)$$

Полученный логоген L_j добавляется в множество $\hat{L}^t$.

Начальным состоянием логогена при обработке ПСЛ является нечеткое множество:

$$CF_0 = \begin{cases} \{x \mid \mu_{a^{t-1}} (-x), x \in CF\}, v_j = 2 \\ \{x \mid 0, x \in CF\}, k_i \neq 2 \end{cases} . \quad (14)$$

По окончании обработки всех морфов по всем позициям текста, получаем нечеткую характеристику текста уровня морфов $\tilde{M}$ (4). Данная характеристика текста указывает расположение в первичном тексте элементов минимальной смысловой информации – морфов. Дальнейший анализ полученной нечеткой характеристики текста позволяет конкретизировать содержащуюся в тексте смысловую информацию, а также анализируя полученную характеристику текста дальше можно выявлять и устранять ошибки, допущенные при наборе текста.

Выводы

Представленные механизмы обработки текстовой информации в виде электронных документов построены с учетом когнитивных особенностей человека, что позволяет создавать модели обработки текстовой информации в природном виде, выделяя смысловые частицы текста. С помощью представленных технологий обработки текста можно также обрабатывать тексты, в которых были допущены ошибки, что соответствует механизмам обработки текста человеком.

Сочетание алгоритмов последовательной и параллельной нечеткой обработки текста дает возможность разделять текстовую информацию на смысловые слои.

Проведенные и описанные в работах [3,6,7,8,10] компьютерные эксперименты подтверждают работоспособность приведенного в данной работе механизма обработки данных.

Литература

1. Новиков Ф.А. Microsoft Word 2003. – БХВ-Петербург, 2004.
2. Костромин В.А. OpenOffice.org - открытый офис для Linux и Windows. – БХВ-Петербург, 2005.
3. Каргин А.А. Ломонос Я.Г. Модель нечеткого текста в интеллектуальной системе терминологической разметки электронных документов. // Вестник Донецкого национального университета. Донецк, 2005г.
4. Солсо Р. Когнитивная психология – СПб.: Питер, 2002.-496с.
5. Шиффман Х. Р. Ощущения и восприятие. – Спб.: Питер, 2003.
6. Ломонос Я.Г. Нечеткая модель терминологической разметки электронных текстов // Вестник ХНТУ. – 2006. – № 1(24). – С. 282–288.
7. Каргин А. А., Парамонов А. И., Ломонос Я. Г. Интеллектуальная система категоризации и интерпретации текстовой информации «Text-Term-Concept». // Збірка наукових праць у чотирьох томах ISDMIT'2006, Том 1 Секція 1, Евпатория – 2006, с. 92-99.
8. Каргин А. А., Ломонос Я. Г. Исследование метода интерпретации аудиальных данных с учетом контекста. // Вестник Херсонского государственного технического университета №1(19), 2004г. Херсон, 2004г. с. 272-277.
9. Люггер Дж.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем // М.: Изд. Дом «Вильямс», 2003.
10. Ломонос Я.Г. Терминологическая разметка текста в автоматизированной системе интеллектуальной обработки текстовой информации // Журнал «Штучний Інтелект» №3'2006 – ІПШІ МОН і НАН України «Наука і освіта», 2006г. – С. 537–547.