Artículo de Investigación. Revista Ciencia e Ingeniería. Vol. 38, No. 3 pp. 291-300, agosto- noviembre, 2017
ISSN 1316-7081. ISSN Elect. 2244-8780. Universidad de los Andes (ULA)

Sobre un modelo de comparación semántica de documentos textuales

About a model of semantic comparison of textual documents

Bermúdez Soto, José Gregorio
Universidad Federal del Sur. ICTIS. Taganrog, Rusia
jbermudesoto@gmail.com

Resumen

En el trabajo se considera un modelo de comparación de documentos textuales para determinar su similitud semántica, limitadoa textos científico- académicos. En baseal análisis de los métodos existentes, se introduceel concepto y método deextracción de "pasajes significativos", el cual garantiza que los segmentos a compararse, tienen un significado semántico completo; se utiliza la presentación de los pasajes significativos en esquemas semánticos, que permiten comparar los elementos de significado de los pasajes; se incorporan las clases semánticas de las palabras en la comparación; y se realizael cálculo de la similitud semánticaentre documentos por los criterios de correctitud y completitud. Se presentan los resultados de los experimentos realizados, junto con su análisis y comparación con otros métodos existentes. La investigación está enmarcadaen las áreas del procesamientoautomático de textos y la lingüística computacional. Deacuerdo con el esquema general del procesamiento del lenguaje natural (PLN), este trabajo se centraen el nivel semántico. La presente investigación y los experimentos que se presentan, fueron desarrollados parael idioma ruso, peroen este documento se presenta su adaptación al idiomaespañol.Los resultados de esta investigación y el modelo propuesto tienen aplicación directa en aplicaciones de detección automática de plagio, para aumentar su efectividad; y en la educación a distancia, para mejorar los métodos de evaluación de respuestas.

Palabras claves: *Similitud textual, comparación de textos, pasajes significativos, presentación en esquemas semánticos, clases semánticas.*

Abstract

In the paper a model of comparing of textual documents is considered to determine their semantic similarity, limited to scientific-academic texts. Based on the analysis ofexisting methods, the concept and method of extraction of "significant passages" is introduced, which guarantees that the segments to be compared, havea complete semantic meaning; we use the presentation of significant passages in semantic schemes, which allow us to compare theelements of meaning of the passages; the semantic classes of words are incorporated in the comparison; and the calculation of the semantic similarity between documents is made by the criteriaof correctness and completeness. We present the results of theexperiments performed, together with their analysis and comparison with other existing methods. The research is framed in theareas of automatic word processing and computational linguistics. According to the general schemeof natural language processing (NLP), this paper focuses on the semantic level. The present research and theexperiments that were presented were developed for the Russian language, but this document presents its adaptation to the Spanish language.The results of this research and the proposed model have direct application in applications of automatic detection of plagiarism, to increase its effectiveness; and in distance education, to improve methods of evaluation of responses.

Keywords: *Textual similarity, text comparison, significant passages, presentation in semantics schemes, semantic classes.*

1 Introducción

En actualidad la búsqueda de la similitude o semejanza entre textos tiene una gran aplicación práctica, incluido la detección de plagio académico y la educación a distancia. En los trabajos (Maurer y col., 2006, Mihalcea y col., 2006) se mencionan tres categorías básicas de detección de similitud textual: comparación basada en palabras; la búsqueda lineal basada en elementos, utilizada por los motores de búsqueda; y el análisis estilístico.

Existen diversos métodos basados en diferentes particularidades de los textos, tal como los métodos basados en la semántica, tanto para la detección de plagio (Bao y col., 2004, Chi-Hong y col., 2007), como para la búsqueda de información (Vishnyakov 2012).

En concordancia con el esquema general de procesamiento de lenguaje natural (PLN), este trabajo se centra en el nivel semántico; pero comprende una descripción general de los pasos y procesos de procesamiento, desde la segmentación, hasta la evaluación final. Los aportes relevantes se comprueban en la introducción del nuevo método de segmentación para la obtención de pasajes significativos, la selección de un método de presentación semántica existente que se corrobora en el trabajo de (Visnyakov 2012); la incorporación de las clases semánticas en la comparación; y la determinación del grado de semejanza en la fase de cálculo de similitud.

En la entrada del proceso de comparación se tienen dos documentos, destinados a la comparación; uno de los cuales se considerará como patrón. En el primer nivel de procesamiento se ejecuta la extracción de pasajes significativos (Bermúdez 2016b). La salida de este primer nivel se convierte en la entrada del próximo nivel, el cual consiste en la presentación semántica de esquemas (grafos) (Bermúdez 2016a, Vishnyakov 2012). Los esquemas de presentación son la entrada para el nivel de detección de similitud semántica entre pasajes, en la que se incorporan las clases semánticas de las palabras; para finalizar con el cálculo de similitud entre los dos documentos.

Encontrar la similitud semánticaentre pares de textos es un problema importante para el PLN. Tal problema surge en varios aspectos de PLN, como la traducción automática, la generación automática de resúmenes, la detección del plagio académico, la evaluación en el campo de la educación a distancia, las pruebas para comprensión de texto, la búsqueda y recuperación de información; y muchos otros, en los cuales es necesario medir el grado de similitud entre dos textos dados.

La búsqueda del grado de similitud semántica de textos ha sido considerada una tarea en muchas conferencias internacionales (Aguirre y col., 2013). Estos aspectos han recibido una considerable atención en los últimos años. Muchos de los modelos desarrollados hacen principalmente énfasis en la búsqueda de características que coincidan en ambos textos, procurando el descubrimiento de significados análogos en ellos.

No obstante a lo dicho anteriormente, los textos de estilo científico-técnico favorecen el procesamiento automático, dada sus características; sin embargo no todas las tareas de PLN para este tipo de textos están totalmente resueltas. Por ejemplo, en los sistemas "anti-plagio", los documentos se comparan para determinar si uno de ellos se escribió exactamente igual que el otro, ya sea en todo o en parte, pero no determinan ninguna coincidencia si el plagiario expone las ideas del autor con otras palabras o paráfrasis.

Los métodos de comparación basados en palabras, la búsqueda lineal basada en elementos, y el análisis estilístico, no proporcionan resultados suficientemente cualitativos, ya que todos los textos tienen una estructura local en diferentes niveles, y para realizar un análisis semántico más preciso, es necesario aplicar métodos que permitan estudiar las estructuras de todos los niveles.

Por lo que se plantea la tarea de proporcionar una detección automática de la similitud semántica entre dos textos comparados mediante la creación de métodos que tengan en cuenta tanto laestructura morfológica del texto como su contenido léxico-semántico.

La mayoría de los investigadores consideran la detección automática de similitudes entre dos textos como tareas independientes sin una conexión entre ellos; y no consideran aspectos del problema tales como, la posibilidad de establecer la similitud del contenido semántico entre dos textos; cuando uno de ellos es plagio del otro, mediante la paráfrasis; o uno es opuesto al otro; o ambos tienen la misma idea, pero no se trata de plagio.

Por ello se propone el desarrollo de un modelo de comparación de textos en lenguaje natural, que permita realizar la extracción de segmentos de texto con un significado completo y revelar la similitud semántica, utilizando algunos componentes desarrollados en estudios previos e introduciendo nuevos métodos de solución, teniendo en cuenta los aspectos semánticos.

Para lograr lo anterior, se requiere implementar la integración de métodos o pasos. Desarrollar un método de segmentación y un método de comparación. Desarrollar algoritmos para la segmentación y comparación. Evaluar la similitud de acuerdo con los criterios de corrección y profundidad. Y realizar experimentos de segmentación y comparación para confirmar la efectividad del modelo.

Nuevas investigaciones sobre el desarrollo de esta propuesta pueden contribuir a métodos para aumentar la eficiencia del procesamiento automático de textos en lenguaje natural, en particular, en la comparación automática de segmentos de texto semánticamente semejantes, escritos con diferentes vocabularios.

2 Los “Pasajes Significativos” como base para la comparación de textos

Para el procesamiento de lenguaje natural siempre es requerido la segmentación del texto, para el consiguiente procesamiento. La segmentación de documentos consiste en

la división automática del documento en partes semánticamente contiguas.

La detección automática de los límites de los segmentos de contenido semántico en el documento es un problema difícil en las tareas de procesamiento de texto en lenguaje natural. Se examinó un conjunto de métodos que intentan resolver este problema, algunos de ellos con buenos resultados, aunque tienen algunas desventajas. Además, muchas de estas soluciones tienen las limitaciones de una aplicación en particular. Se realizó entonces el análisis de algoritmos existentes para la segmentación de documentos, con el fin de revelar sus ventajas y desventajas, así como su utilidad para resolver la tarea. Entre los trabajos, algoritmos e investigaciones evaluadas se encuentran los métodos básicos basados en palabras, el método de N-gramas, la extracción de pasajes textuales arbitrarios, la segmentación en sub-temas, en particular el algoritmo "TextTiling" y otros. (Hearst 1997, Salton 1989, Jurafsky col., 2008, Silva y col.,1999, Kaszkiel y col., 2001, Heinonen 1998).

La idea principal del método propuesto de segmentación, se basa en la necesidad de extraer pasajes textuales lo más cortos posible, pero que contengan un significado completo. Es decir, que se requiere obtener un segmento de texto lo más pequeño posible con un significado completo. El método de extracción de pasajes arbitrarios posibilita extraer segmentos de textos que, por su tamaño, facilitan el procesamiento a niveles superiores. Pero, no hay garantía, cuando se divide un texto, que un pasaje de texto arbitrario tendrá algún significado semántico.

Se propuso entonces un método que posibilita extraer pasajes significativos de textos, basado en la identificación de los verbos conjugados y las anáforas.

Para la extracción de pasajes, se requiere el proceso de segmentación en palabras, con el fin de determinar el papel gramatical de cada palabra e identificar qué palabras funcionan como referencias anafóricas. También implica la identificación de los verbos en el segmento, para asegurar que estos segmentos contienen un significado completo.

En este contexto, las palabras que se consideran enlaces anafóricos son los pronombres y adverbios con la función gramatical de enlaces anafóricos: pronombres personales (él, nosotros, vosotros, ellos, etc.), los pronombres relativos (quién, qué, cómo, etc.); y pronombres demostrativos (este, que, como, etc.). Mientras que los verbos los llamaremos Tipo A: en las formas personales de los modos indicativos, subjuntivos e imperativo.

Procedimiento:

1. Dividir el texto en palabras y marcar cada palabra de acuerdo con su rol gramatical. Estos algoritmos existen completamente implementados y son conocidos como: Tokenización y Pos-Tagging.
2. Incluir palabras en el segmento de texto, hasta que se hallan incluido: 1) un verbo del Tipo "A"; 2) todos los elementos que representan anáforas que estén a la derecha del verbo y con el antecedente respectivo antes del siguiente signo de puntuación del tipo "." ó ";"; 3) el subsiguiente signo de puntuación del tipo "." ó ";".
3. Repetir el paso 2 hasta que las palabras terminen.

La ventaja de este algoritmo es que este enfoque proporciona en cada segmento un alto grado de cohesión léxica. Esta es una propiedad importante del texto, ya que los bloques de texto que están relacionados por anáforas generalmente representan un segmento que incluye el significado completo, y estos no son demasiado largos. Además, aunque los signos de puntuación se emplean para referirse a los límites de un segmento, en sí mismos no son un criterio de parada.

Por lo tanto, un pasaje significativo se define como *"un conjunto de frases u oraciones consecutivas en un documento en el que no hay referencias anafóricas asociadas con las palabras de otro segmento y en las que hay al menos un verbo cuyo tipo y categoría expresa acción"*. El tamaño de cada pasaje se mide por el número de frases u oraciones que lo forman, este parámetro se determina por la propia redacción del documento.

Como concepto, la palabra significado se refiere al contenido mental que le es asignado a un término (palabra) en cuanto a la lingüística. En otras palabras, es el concepto o idea que se asocia al término (Márquez 2008). Sin entrar en profundidad, respecto a la interpretación semántica de las oraciones; para los propósitos de este estudio, diremos que un pasaje es significativo, cuando en el segmento no haya referencias anafóricas asociadas a palabras de otro segmento; y que tiene al menos un verbo, del tipo y categoría que expresa acción.

El pasaje significativo como unidad de procesamiento de texto, tiene varias ventajas sobre el párrafo, oraciones, frases o palabras. El concepto formal de pasaje significativo se presenta en el siguiente aparte, a los efectos de describirlo junto con la presentación formal de comparación textual.

3 Método de comparación basado en pasajes significativos y cálculo de similitudes entre documentos.

En general los métodos de comparación, mayormente utilizados para la búsqueda de información, utilizan diferentes perspectivas para solucionar el inconveniente de reconocer los documentos que cumplan con los requerimientos de información del usuario. De manera general según (Llopis 2003), si se tienen un documento *D* y una consulta *Q*, la idea final es medir la semejanza o relevancia entre los dos:

$$sem(D,Q)=? \quad (1)$$

Con el objeto de determinar dicha relevancia, los sistemas de búsqueda y recuperación de información aplican un conjunto de funciones, llamadas medidas de similitud, que cuantifican ese valor de relevancia entre el documento y la consulta. Principalmente, estas medidas se basan en la cantidad de palabras que tienen tanto el documento como la consulta (ob. Cit 2003).

Para el método propuesto, tendremos entonces que el objetivo final es medir la similitud o relevancia entre dos textos D_1y D_2:

$sem(D_1,D_2)$=? (2)

Además se debe realizar un conjunto de pasos adicionales:

1. Descomponer los documentos D_1y D_2 en pasajes significativos respectivamente

$D_1 \rightarrow P_{11}, P_{12}, P_{13}, ...P_{1n}$ (3)

$D_2 \rightarrow P_{21}, P_{22}, P_{23}, ...P_{2m}$ (4)

Donde P_{1i} y P_{2j} son pasajes significativos de texto de los documentos D_1 y D_2 respectivamente.

2. Representar los pasajes significativos en esquemas semánticos

$\forall\, i \in 1,...n \rightarrow P_{1i} := \hat{p_{1i}}$ (5)

$\forall\, i \in 1,...m \rightarrow P_{1j} := \hat{p_{1j}}$ (6)

donde pˆes el esquema semántico del pasaje.

3. Calcular la similitud X_{ij} de los pasajes de D_1 con respecto a los pasajes de D_2

$\forall\, ij \in 1,...n;\ 1,...m \rightarrow sem(P_{1i}, P_{2j})=X_{ij}$ (7)

4. Calcular la similitud entre los documentos D_1 y D_2 en función de la similitud entre los pasajes que los conforman

$sem(D_1,D_2)= f(X_{ij} .. X_{nm})$ (8)

Como se puede observar, el modelo propuesto tiene una complejidad mayor que la de los métodos básicos, dado que debe realizar un número de tareas adicionales. No obstante, el método propuesto ofrece ventajas, que compensan el incremento de complejidad con un incremento de la efectividad en los resultados obtenidos.

Formalmente, la definición de pasajes de texto significativos en el método propuesto es la siguiente (ob. cit., 2003):

- Sea *D*un documento formado por *N* frases fj.

$D = (fj, ...fN)$ (9)

- Se definen los pasajes significativos P_i, del documento *D*, de la siguiente manera:

$P_i = (fj, ...fq)$; ... $P_n = (fk, ...fN)$; (10)

donde: *n* es el número de pasajes resultantes de la segmentación.

Pero tales pasajes textuales no son arbitrarios, sino que están condicionados por el criterio de parada de segmentación indicado anteriormente. Así entonces en todos los pasajes P_{i+1} no existen referentes anafóricos relacionados con antecedentes en P_i; al tiempo que en cada pasaje P_i existe al menos un verbo del tipo *A*.

Por tanto, sean *a* un verbo del tipo indicado anteriormente, *h* un elemento anafórico cualquiera, *c* su respectivo antecedente; y el símbolo ":=" la relación inequívoca que indica que *h* representa a *c* (*h*:=*c*); entonces la definición de los pasajes significativos P_i quedará:

$P_i = (fj, ...fq), / \forall_i ((\exists_{a \in} P_i) \wedge (\nexists_{h \in} P_{i+1} / (h{:=}c) \wedge c \in P_i))$(11)

Para la determinación de la semejanza entre dos documentos, debe solucionarse previamente la similitud entre las partes significativas que lo conforman; y en nuestro caso además se debe abordar el punto de la representación de los pasajes significativos en esquemas semánticos.

No obstante a lo indicado anteriormente, existen también métodos que calculan la similitud entre dos documentos a través de algoritmos basados en la frecuencia de ocurrencia de términos entre ambos documentos, pero esto no es más que un método que extiende la comparación entre una consulta y un documento, en el que uno de los documentos es tratado como consulta. De tal manera que basta con examinar el cálculo de similitud entre consulta y documento. Pero antes de abordar los aspectos del cálculo de similitud, en nuestro caso, como ya se advirtió, trataremos brevemente el aspecto de la representación en esquemas semánticos.

La representación de texto en esquemas, es una tarea a nivel semántico dentro del PLN, que se viene utilizando, por lo general en representaciones de redes semánticas de presentación del conocimiento (Quillian, 1968). A partir de esto se obtuvieron conceptos y perspectivas de procesamiento tales como estructuras de datos, redes semánticas, grafos conceptuales, grafos de marcos y otros.

La presente investigación no trata sobre el procesamiento del conocimiento; sino más bien del uso de las estructuras de grafos para la presentación, como una ventaja para el cálculo de la similitud textual.

En este sentido este trabajo, en cuanto al punto de la presentación de los pasajes textuales significativos, se basa totalmente, en el método y enfoque presentado por (Vishnyakov 2012); quien utilizando dichos esquemas de representación semántica, introduce la definición de funcionalidad sentido expresiva, la cual utiliza para procesar fragmentos de textos de una partición cualquiera y comparar su sentido de expresión con otro fragmento para revelar su proximidad semántica (figura 1).

En el esquema de presentación de Vishnyakov, se parte de un segmento de texto cualquiera; perfectamente puede tratarse de un pasaje significativo; para construir el árbol de dependencias, utilizando la regla: si dos palabras de un fragmento están conectadas por una dependencia directa, entonces en el árbol, la principal de ellas se colocará más alto y la dependiente más abajo. Desde la palabra principal se traza un arco hacia la palabra dependiente. Se realizan tales acciones para todas las palabras del fragmento, resultando en el árbol de dependencias como en la figura 1.

En la escritura y composición de textos, invariablemente sucede que una palabra dependiente normalmente precisa y caracteriza el sentido de la palabra principal. En este caso, el orden de las palabras en la frase puede ser diferente, y la relación de las palabras principales y dependientes puede establecerse por sentido y gramaticalmente. En este método el autor establece una conexión en el significado, que se determina mediante las preguntas que se colocan desde la palabra principal a la palabra dependiente.

Posteriormente, para la construcción del esquema semántico de presentación del fragmento, partiendo del árbol de dependencia se procede: de izquierda a derecha, comenzando desde el nodo más a la izquierda que no tenga hijos (hojas del árbol), construyendo los elementos de significado de cada par de palabras (dependiente a principal),

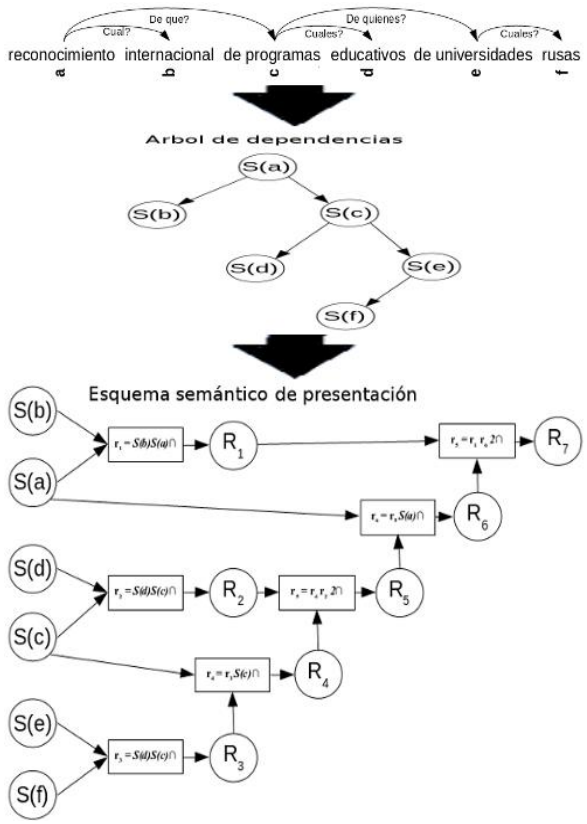

Fig. 1. Esquema de presentación semántica

Esto para el primer nivel (primera interación), hasta que se agoten; después se eliminan del árbol los nodos sin hijos, para repetir el paso de construcción del siguiente nivel de elementos de significado, se repiten ambos pasos hasta que se agoten todos los nodos. Nótese que en algunos casos al construir los elementos de significado puede haber una doble, triple y hasta N intersecciones de una misma palabra, tal es el caso de R_5 y R_7 en el ejemplo. Ver figura 1.

Lo más importante en este tipo de esquema, es que los nodos construidos no son sólo palabras, sino desde frases (primer nivel) hasta el propio segmento completo, como resultará siempre con el último elemento de significado. Precisamente estos elementos de significados serán comparados para determinar la similitud entre un par de pasajes significativos, correspondientes a los textos a comparar.

En el ejemplo de la figura 1, los elementos de significado quedarán así: R_1=reconocimiento internacional; R_2=programas educativos; R_3=universidades rusas; R_4=programas de universidades rusas; R_5= programas educativos de universidades rusas; R_6= reconocimiento de programas educativos de universidades rusas; R_7= reconocimiento internacional de programas educativos de universidades rusas.

La medida de similitud permite cuantificar la semejanza entre dos segmentos de texto (ya sea un documento completo o un pasaje del mismo) y una consulta; o en nuestro caso entre dos pasajes significativos. Tradicionalmente estas medidas se basan fundamentalmente en los términos que comparten el texto y la consulta así como en la importancia discriminatoria de cada término (Llopis 2003).

Los métodos de búsqueda y recuperación de información realizan estos cálculos de similitud, definiendo un documento D como un conjunto de pares de valores (d_i, n_i), en los cuales d_i sería el término n_i el número de veces que aparece dicho término en el documento. El valor N representa el tamaño del documento, en cuanto al número de términos diferentes que lo forman; así entonces:

$$D = ((d_1, n_1), (d_2, n_2), \ldots (d_N, n_N)) \quad (12)$$

Por otra parte, en el mismo enfoque de presentación, la consulta Q, se define como un conjunto de pares de valores (q_i, m_i), en los cuales q_i sería el término y m_i el número de veces que aparece dicho término en la pregunta. el valor K indica el número de términos diferentes que forman la consulta, así entonces:

$$Q = ((q_1, m_1), (q_2, m_2), \ldots (q_K, m_K)) \quad (13)$$

La medida de similitud entre Q y D se calcula, entre otros métodos, en función de (ob. cit., 2003):

- El número de palabras que existen tanto en la consulta como en el documento.
- El número de veces que aparecen en ambos (consulta y documento), dichas palabras.
- El peso x_i de la palabra dentro de la colección de documentos. Este peso x_i de una palabra t_1, se define en función del número de documentos de la colección en los que aparece dicha palabra.

Así, la medida de semejanza se define de la forma:

$$sem(D, Q) = Y\forall_{i \in Q \wedge D} (t_i, n_i, m_i, x_i, N) \quad (14)$$

donde: Y define un método para cuantificar el valor de la semejanza entre documento y consulta, en función de los parámetros.

Según (ob. cit 2003) hay otros métodos que utilizan pasajes como unidad de procesamiento, el cálculo de similitud entre pasaje y consulta es igual, pero sustituyendo las apariciones de documento por las de pasaje, para luego calcular la similitud entre la consulta y el documento en función de la similitud de todos los pasajes. Además, no existe en muchos de ellos una asignación directa del modelo que define la forma de segmentación del documento en pasajes y la medida de similitud utilizada.

El planteamiento que se hace en esta investigación es diferente. En primer lugar cabe recordar que los pasajes significativos son unidades completas con un significado intrínseco, cuyo tamaño queda determinado por la propia redacción del documento. Y que los pasajes significativos, se representan en esquemas semánticos para la comparación. Por otra parte el presente modelo incluye las comparaciones en función de las clases semánticas y no sólo por la igualdad exacta de las palabras.

En cuanto a las clases semánticas, en este trabajo, se considera como tal a un conjunto de términos, palabras o expresiones, las cuales poseen un significado similar. En un sentido estricto, las clases semánticas en términos de la lingüística es una asociación de palabras cuyos valores

corresponden a un idéntico fenómeno o concepto de la realidad (Rodríguez 2004).

Por otro lado en la lingüística se emplea también el concepto de campo semántico, que se refiere a un conjunto de palabras o elementos significantes con contenidos relacionados, debido a que comparten un núcleo de significación o rasgo semántico común y se diferencian por otra serie de rasgos semánticos que permiten hacer distinciones (ob. Cit 2004).

Pero en la presente investigación se hace referencia a la clase semántica, como un conjunto de palabras (términos), que tienen una relación de significado similar o disímil; ya sea que, en el sentido estricto de estos conceptos tradicionales, pertenecen al mismo campo semántico, a la misma clase, a ambos o a ninguno.

En tal sentido para esta investigación se define como clase semántica al conjunto de términos, palabras o expresiones, que tienen una relación de significado entre sí, tal que el intercambio de una por otra en el contexto de un texto escrito, no altera en modo alguno el significado de la frase, oración o pasaje textual.

Otra diferencia significativa del método que se propone en esta investigación, lo comprende el que se considere uno de los textos a comparar como un patrón, lo que permite establecer predeterminadamente ciertas condiciones.

En particular en el patrón se requiere que de forma predeterminada, con la asistencia humana, se realice la determinación del grado de similitud, de palabras que pertenecen a la misma clase semántica. Esto significa que de alguna manera, al experto o usuario, se le presenta una palabra determinada junto con una lista de palabras similares, que pueden reemplazar a la primera en el texto y evaluar el grado de similitud que estas tienen, en función del sentido semántico que refleja en el texto.

De lo anterior se desprende que para nuestro caso cada pasaje significativo del texto patrón, se convierte en un conjunto de frases, cuyos términos tienen asociados una lista de palabras cada uno.

Así entonces sea fj una frase parte de un pasaje significativo $P_i = (fj, ...fq)$; y si $fj = (tk, ...tm)$, donde cada tk, es un término de la frase; la representación del pasaje textual indicado en (11), quedará:

$$P_i = (tk, ...tq), / \forall_i ((\exists_a \in P_i) \wedge (\nexists_h \in P_{i+1} / (h{:=}c) \wedge c \in P_i)) \quad (15)$$

Y los pasajes textuales del texto patrón, será:

$$\underline{P_i} = (Tk, ...Tq), / \forall_i ((\exists_a \in \underline{P_i}) \wedge (\nexists_h \in \underline{P_{i+1}} / (h{:=}c) \wedge c \in \underline{P_i})) \quad (16)$$

donde: cada término Tk, tienen asociado una lista de palabras y un peso asociado a cada palabra, en la forma:

$$Tk = (t_1, X_1), (t_2, X_2), \ldots (t_N, X_N) \quad (17)$$

La similitud entre dos pasajes $\underline{P}_1$, P_2 dependerá de las apariciones de los términos del P_2 en los términos de $\underline{P}_1$, y sus pesos asociados, dados en (17), de la forma:

$$sim(\underline{P}_1, P_2) = \Phi \forall_{ij \in} P_1 \wedge P_2 (Ti, t_i, X_i, t_j, N) \quad (18)$$

donde: Φ define el método para cuantificar el valor de la similitud entre los pasajes, en función de los parámetros, en la forma:

$$\Phi_{\text{semántica/clases}} = \frac{\sum_i^k \frac{\sum p_j}{l}}{n} \quad (19)$$

donde: p es el factor de coincidencia entre las palabras que participan en la comparación, para cada elemento de significado, según la clase semántica en el intervalo [0,1], $p = 1$, si la palabra es idéntica, $p = 0$ si la palabra no está en la clase semántica; y $p = (0,1)$ en dependencia con el grado de sinonimia; l es la cantidad de palabras de cada elemento de significado; k es la cantidad de elementos de significado del pasaje del texto a comparar; y n es la cantidad general de elementos de significado del pasaje del texto patrón.

Como ya se ha indicado en la expresión (14), la mayoría de los métodos de recuperación de información calculan la similitud del documento en función de la similitud de sus pasajes, en los que la función Y puede ser fundamentalmente, la del pasaje de mayor similitud o la suma de similitudes.

En nuestro caso la situación es diferente, dado el objetivo de la comparación, trata de dos documentos, como se indicó en (8). De tal manera que en nuestro caso, para determinar la similitud entre los documento, se emplearan a grandes rasgos ambos enfoques.

Lo anterior se debe principalmente, a que al comparar un texto con otro, cada pasaje significativo del texto a comparar, se debe confrontar con todos los pasajes del texto patrón en una relación n:m; de lo cual se escogerá el de mayor similitud, en la forma:

$$sem(\underline{P}_1, P_2) = max_{\forall ij \in P1 \wedge P2} sem(P_{1i}, P_{2j}) \quad (20)$$

Siendo que dicho valor máximo de similitud descarta tanto al pasaje del texto comparado, como al pasaje del texto patrón, que participaron en dicha comparación.

La determinación de la correctitud y completitud dependen directamente del objetivo de la comparación y de su consecuente evaluación. Un criterio viable que surge de los resultados obtenidos en la etapa anterior, es la correctitud, pero ahora con respecto a todo el texto, es decir, el coeficiente de correctitud C determinado por la fórmula:

$$C = \frac{\sum_1^q \Phi_i}{m}, \quad (21)$$

Donde Φ es el resultado obtenido para cada comparación de la etapa anterior; q es la cantidad de pasajes del texto a comparar; y m es la cantidad general de pasajes del texto patrón.

La completitud puede ser obtenida de la proporción simple de la cantidad de pasajes del texto a comparar entre la cantidad de pasajes del texto patrón, es decir que la completitud S, se determina según la formula:

$$S = \frac{q}{m}, \quad (22)$$

Al tiempo se puede evaluar el resultado final con el promedio de los dos coeficientes obtenidos anteriormente; es decir que R, se determina por la fórmula:

$$R = \frac{C+S}{2}, \quad (23)$$

4 Integración de métodos y algoritmos para la comparación textual

La solución planteada a través del modelo propuesto, se basa en el esquema general de PLN, pero se completa con un conjunto de pasos específicos, modificando métodos individuales del esquema y organizando relaciones adicionales entre ellos. De manera tal que las sub tareas requeridas para determinar la similitud semántica de los textos comparados son las siguientes:

1. Extracción de pasajes significativos.
2. Presentación de los pasajes en esquemas semánticos.
3. Determinación del grado de semejanza semántica entre pasajes, de acuerdo a las clases semánticas.
4. Determinación de correctitud y completitud del texto en comparación con el patrón.

El modelo combina varios elementos y métodos y/o algoritmos existentes; con modificaciones de algunos métodos, teniendo en cuenta las características del problema. Este enfoque proporciona una solución efectiva que se presenta esquemáticamente en la figura 2.

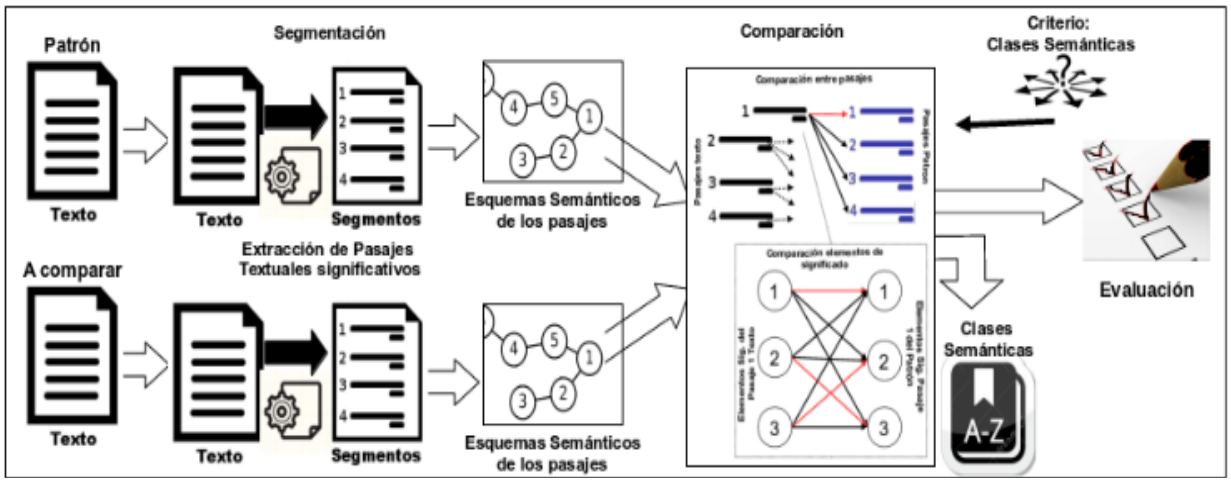

Fig. 2. Modelo de integración de comparación textual

5 Resultado de experimentos de segmentación y comparación textual.

Se compararon algunos de los métodos de segmentación analizados con el método de extracción de pasajes significativos y una segmentación realizada por personas; en particular se compararon los métodos: extracción de pasajes arbitrarios y el método "TextTiling". En el experimento para la segmentación de pasajes arbitrarios se utilizó el punto como criterio de parada. Para el método "TextTiling" (Hearts, 1997), se utilizaron los valores de w=5 y k=2 respectivamente.

Para realizar el experimento, se utilizaron cien (100) textos, contentivos de las introducciones de artículos científicos de la colección de publicaciones de la Cátedra de Análisis de Sistemas y Telecomunicaciones del Instituto de Tecnologías de Computación y Seguridad Informática de la Universidad Federal del Sur, cada uno de aproximadamente una página. Los cuales fueron procesados como se indica a continuación para el primer texto: titulado "Enfoque hacia la definición de meta-sistema como sistema" (Rogosov, 2013), el cual consta de 3 párrafos, 214 palabras.

La segmentación manual basada en el juicio humano fue realizada por 10 personas por cada texto, para el texto 1 se escogieron como válidos los 10 límites de segmento (16, 31, 43, 57, 72, 110, 122, 139, 164 y 190.), donde al menos existieron 6 coincidencias (figura 3).

Se calcularon los valores de la métrica "*WindowDif*" para las segmentaciones de los tres algoritmos en los cien textos. La métrica "*WindowDif*" (Pevzner y Hearst, 2002) emplea una ventana corrediza de tamaño *k* para un recorrido por todo el texto y revelar las desigualdades entre la segmentación de referencia y la que se está evaluando. Donde, *k* se obtiene de la mitad del promedio del tamaño que tienen los segmentos en la segmentación de referencia. Conforme a como la posición de la ventana va avanzando, se va determinando, para las dos segmentaciones (referencia y evaluada), la cantidad de límites que se encuentran en dicha ventana, cuando el número de límites no es el mismo, el algoritmo que se evalúa recibe una penalización. Al final se totaliza dichas penalizaciones para todo el texto y se pondera este valor en el intervalo [0,1]. así entonces la métrica "*WindowDif*" alcanza el valor de 0 si el algoritmo asigna todos y cada uno de los límites correctamente y es 1 si por el contrario es incorrecto para todos los casos.

Se compararon los resultados obtenidos de la segmentación manual y las realizadas con los algoritmos estudiados: extracción de pasajes arbitrarios, "TextTiling" y extracción de pasajes significativos.

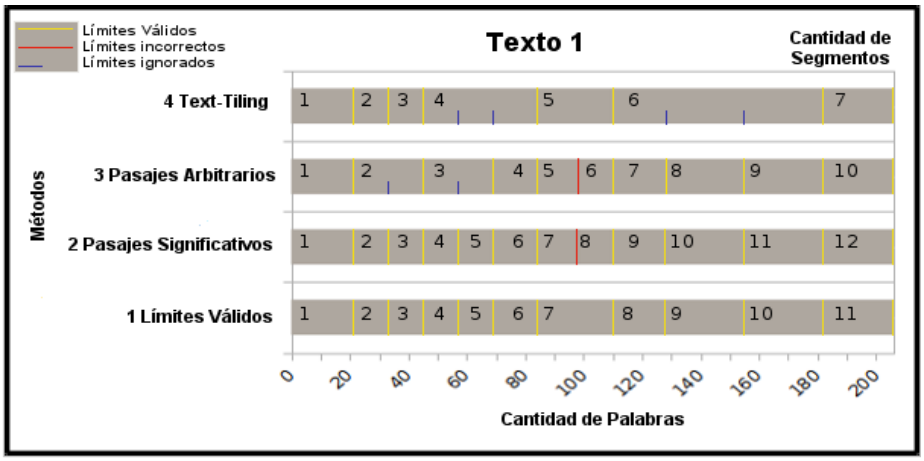

Fig.3. Comparación de los métodossegmentación para texto 1

Los resultados de los algoritmos coinciden en ocasiones con los límites que se especifican como válidos para los cien textos de referencia. Para el caso del método propuesto, los límites son los más próximos a los válidos, como puede apreciarse en las figuras 3 y 4. Los resultados obtenidos por la métrica "*WindowDif*", se presentan en la tabla 1; en cuyo caso el método propuesto tiene para el promedio de los cien textos el menor valor de dicha métrica, que implica más proximidad con los límites válidos.

Tabla 1. Valores de la métrica "*WindowDif*"

Algoritmo	**"*WindowDif*"**
Pasajes significativos	0,046
Pasajes arbitrarios	0,161
"TextTiling"	0,232

Para la determinación de semejanza, se llevó a cabo una comparación de algunos métodos existentes con el método propuesto en esta investigación y un análisis realizado por personas. En particular se utilizaron los siguientes métodos y programas existentes para la comparación:

1. Los métodos de comparación de textos basados en el coeficiente de similitud de Jaccard, la similitud coseno y la distancia de Levenshtein, utilizando para ello un programa online de algoritmos de similitud entre cadenas de texto, basado en el lenguaje de programación php (Francesc 2015).
2. El método de análisis de semántica latente y otros métodos de búsqueda y recuperación de información utilizando para ello el programa online de detección de plagio "plagiarisma.net"; el cual está basado en la utilización de los motores de búsquedas "Google", "Babylon" y "Yahoo".
3. El programa de detección de plagio de la Universidad Federal del Sur (UFS) denominado "Anti-Plagio", el cual se supone basado en el método de búsqueda por análisis de semántica latente y otros algoritmos propios de la empresa propietaria del software.
4. El método de determinación de similitud para la recuperación de información indicado en el trabajo (Vishnyakov 2012), que llamaremos "*Φ* - semántica"

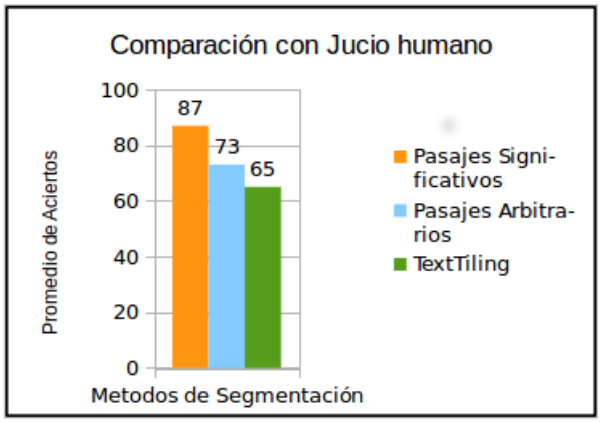

Fig. 4. Comparación de los métodos para 100 textos analizados

Para el experimento fueron utilizados cuatrocientos (400) textos, a saber: 1) cien (100) textos originales, contentivos de las introducciones de artículos científicos de la colección de publicaciones que se mencionan en el experimento anterior; 2) Cien (100) textos, contentivos de plagio, obtenidos a partir de la modificación de los originales, los cuales fueron escritos intencionalmente, sustituyendo en el texto original, algunas palabras y frases similares; 3) Cien (100) textos opuestos a los textos originales, los cuales fueron escritos intencionalmente; y 4) Cien (100) textos de las interpretaciones de los textos originales, los cuales fueron escritos intencionalmente, como respuesta a una pregunta sobre el contenido general del texto.

Para los algoritmos de similitud entre cadenas de texto y los sistemas de detección de plagio, se compararon los trescientos textos de los tres tipos con relación a los cien textos patrones, incluyendo la comparación con si mismo como evaluación de control, estos dan como resultado un porcentaje de similitud entre los textos ingresados.

Igualmente para el método de Vishnyakov y el método propuesto en este trabajo, se realizaron las comparaciones textuales de los textos patrón con los tres textos tipos.

Para el método propuesto en este trabajo se consultaron cien (100) estudiantes del área de tecnologías de información y sistemas, a quienes se les presentó cada palabra o frase de los textos patrón junto con una lista de cinco sinónimos posibles y no más de dos antónimos o frases contrarias. Se les solicitó otorgar el grado de semejanza de dichas palabras que pertenecen a la misma clase semántica, las cuales fueron seleccionadas de WordNet para el idioma ruso. Para los antónimos o frases contrarias se les solicitó su verificación, considerándose válidos los que obtuvieron más de 60 % de aceptación. Los resultados promedios obtenidos para cada palabra de la clase semántica se consideraron como su grado de semejanza.

Cien (100) encuestados del área de las tecnología de la información analizaron cuatro textos cada uno, se les indicó que el texto número uno era un texto original en comparación con los otros tres. Se les pidió que estudiaran a fondo cada texto para obtener respuestas a las preguntas sobre similitud y plagio, todo en relación con el significado expresado en el texto.

Las variantes de las respuestas se presentaron en la escala cualitativa de Likert. Los resultados cualitativos se convirtieron en cuantitativos en una escala porcentual, para compararlos con los resultados de los métodos analizados, tomando como referencia los resultados del análisis de expertos. Los resultados anteriores y su comparación con los métodos utilizados y el método propuesto se presentan y analizan a continuación(ver figura 5).

En promedio, en cuanto al nivel de similitud: el 91% indicó que los cien textos tipo plagio, relativos a los originales, es similar o muy similar. El 83% indicó que los cien textos eran significativamente opuestos o completamente opuestos a los originales. Mientras que el 75% ha confirmado que las cien respuestas eran similares o similares en un pequeño grado; que se traduce en porcentajes de similitud de esta manera: textos plagio = 84%; textos opuestos = 82% y textos respuestas = 42%.

Los resultados anteriores se compararon con los resultados de otros métodos indicados en la figura 5. Como puede verse, el método propuesto para los tres grupos de textos (plagio, opuestos, respuestas) tiene el valor más aproximado con respecto a las opiniones de los encuestados, incluso para los textos del grupo tipo plagio, mientras que otros métodos dan resultados distantes o no determinan similitudes. Mención especial son los resultados obtenidos y presentados para el algoritmo de distancia de Levenshtein, que tiene una característica especial, si los textos se introducen con algunos cambios en el orden de los párrafos en relación con los fragmentos, los resultados se reducen significativamente, mientras que otros algoritmos y métodos retienen el mismo porcentaje; esto se debe al hecho de que el algoritmo de distancia de Levenshtein es el número mínimo de operaciones necesarias para transformar una cadena de caracteres en otra y, con un cambio en el orden de los párrafos, aumenta el número de operaciones. Pero cambiar el orden de los párrafos de un texto no cambia su significado, y más aún no puede disimular el plagio, en este sentido, este algoritmo es ineficaz para fines comparativos.

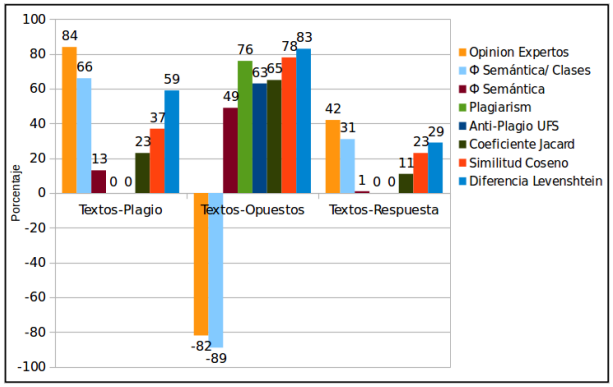

Fig. 5. Resultados métodos de comparación 100 textos analizados

Es importante mencionar que en el caso del grupo de los textos opuestos, los encuestados apuntan al valor opuesto en relación con el original, el método propuesto determina la semejanza de un valor negativo, mientras que los métodos comparados detectan similitud.

6 Conclusiones

En el presente trabajo se propone y desarrolla un modelo compuesto por los siguientes pasos y métodos: Extracción de pasajes significativos; presentación de los pasajes en esquemas semánticos; determinación del grado de semejanza semántica entre pasajes, de acuerdo a las clases semánticas; y determinación de correctitud y completitud del texto en comparación con el patrón.

Además se muestra que el método de extracción de pasajes significativos, permite dividir el texto en una lista de pasajes que transmiten un significado completo, en contraste con los métodos existentes, que se basan en signos de puntuación o criterios estadísticos, que no garantiza que el pasaje de texto tendrá un significado completo.

El método propuesto de comparación semántica entre pasajes significativos, con el uso de las clases semánticas, permite comparar dos textos que transmiten el mismo sentido o el sentido opuesto, cuando se escriben sustituyendo parte del vocabulario, a diferencia de los métodos existentes que sólo miden el valor máximo de similitud.

El método de extracción de pasajes significativos y el método de comparación con el uso de las clases semánticas, fueron probados en experimentos con cien textos de estilo científico-académico; lográndose demostrar que los mismos tienen mayor efectividad que los métodos analizados, para obtener segmentos de texto con significado completo y detectar similitud semántica en caso de sustitución de vocabulario.

Referencias

Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Weiwei Guo, 2013, *Sem-2013 shared task: Semantic textual similarity. In 2nd Joint Conference on Lexical and Computational Semantics (*SeM), Atlanta pp. 32–43.

Bao JP, Shen JY, Liu XD, Liu HY, Zhang XD,2004, Semantic Sequence Kin: A Method of Document Copy Detection. In Proceedings of advances In Knowledge Discovery and Data Mining, vol. 3056, Sydney, pp. 529-538.

Bermúdez S. J. 2016a. Бермудес С. Х. Г. Enfoque hacia la creación de un modelo de comparación semántica de textos. Подход к созданию модели семантического сравнения текстов. Revista "Информатизация и связь" vol. 2-2016, pp. 121-126. Moscú.

Bermúdez S. J. 2016b. Бермудес С. Х. Г. Sobre un método de extracción de pasajes como base para la comparación textual. О методе извлечения значимых текстовых пассажей как базы для текстового сравнения. Revista "Информатизация и связь" vol. 3-2016, pp. 147-153. Moscú.

Chi-Hong, L. Y Yuen-Yan, C. 2007. A Natural Language Processing approach to automatic Plagiarism Detection. In Proceedings of the 8th a CM Conference on Information Technology education (SIGITe'07), pp. 213–218. Florida.

Francesc Ll. C. 2015. Algoritmos de similitud entre cadenas de texto (php). URL: http://francescllorens.eu/00tokenizer/dst.php.

Hearst, Marti A. 1997. TextTiling: segmentating text in to multi-paragraph subtopic passages. Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=972687.

Heinonen O. 1998. Optimal Multi-Paragraph Text Segmentation by Dynamic Programming. Helsinki: University of Helsinki. URL: http://www.aclweb.org/anthology/C98-2239.

Jurafsky, D. and Martin, J.H. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. URL: http://www.deepsky.com/~merovech/voynich/voynich_manchu_reference_materials/PDFs/jurafsky_martin.pdf

Kaszkiel M. y Zobel J. 2001. Effective Ranking with Arbitrary Passages», Journal of the american Society, for Information Science (JASIS). URL: https://pdfs.semanticscholar.org/64fc/fa996acd5f0c5540a161c359fc343601cdac.pdf.

Llopis F. 2003. Un sistema de recuperación de información basado en pasajes. Tesis Doctoral. Universidad de Alicante. Alicante.

Márquez D. N. 2008. Formalización del significado a tráves de la anáfora pronominal: Una introducción a la lógica de predicados. Tesis de Lingüista. Universidad Nacional de Colombia. Bogota.

Maurer, H., Kappe, F. y Zaka, B. 2006. Plagiarism - a Survey. Journal of Universal Computer Science, 12 (8), 1050-1084.

Mihalcea R., Corley C. and Strapparava C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the 21st National Conference on artificial Intelligence, pp.775–780.

Pevzner L., Hearst M. 2002. A Critique and Improvement of an evaluation Metric for Text Segmentation. Computational Linguistics. URL: http://people.ischool.berkeley.edu/~hearst/papers/pevzner-01.pdf.

Quillian R. 1968. Semantic Memory, in M. Minsky (ed.), Semantic Information Processing.

Rodríguez J. 2004. Análisis estructural y significado lingüístico. Revista "Filosofía y Lingüística", vol. 30, pp. 181-203. Costa Rica. URL: https://revistas.ucr.ac.cr/index.php/filyling/article/viewFile/4461/4278.

Rogozov Y. I. 2013. Рогозов, Ю. И. Enfoque hacia la definición de meta-sistema como sistema. Подход к определению метасистемы как системы.Revista Труды Института системного анализа РАН. Nº 4. pp. 92-110. URL: isa.ru/proceedings/images/documents/2013-63-4/t-4-13_92-110.pdf.

Salton G. 1989. Automatic Text Processing : The Transformation, analysis, and Retrieval of Informationby Computer.

Silva, J., y Lopes, G. 1999. A local Maxima Method and a Fair Dispersion Normalization for extracting Multiword Units. In: Proceedings of the 6th Meeting on the Mathematics of Language. URL: http://hlt.di.fct.unl.pt/jfs/MoL99.pdf.

Vishnyakov R. Y. 2012. ВишняковР. Ю. Desarrollo e investigación de la presentación formal y esquemas semánticos de textos de estilo cientifico-técnico para el mejoramiento de la efectividad de la búsqueda de información. Разработка и исследование формализованных представлений и семантических схем предложений текстов научно-технического стиля для повышения эффективности информационного поиска. Tesisdoctoral. Диссертация на соискание ученой степени кандидата технических наук. Южного федерального университета. Таганрог. Universidad Federal del Sur. Taganrog.

Recibido: *7 Diciembre de 2016*

Aceptado: *27 de Julio de 2017*

Bermúdez, José: *Lic. Ciencias y Artes Militares, Ingeniero de Sistemas, Doctorando en Ciencias de la computación e ingeniería en el Instututo de Tecnologías de Computación y Seguridad de la Información de la Universidad Federal del Sur, Taganrog-Rusia.*