Juan Bautista Llorens Morillo, Jorge Morato Lara, José Antonio Moreiro González, Manuel Velasco. Características textuales como medida cualitativa de la información en la generación semiautomática de tesauros. Procesamiento del lenguaje natural, 1998, Nº. 23, pp. 61-68.

CARACTERISTICAS TEXTUALES COMO MEDIDA CUALITATIVA DE LA INFORMACIÓN EN LA GENERACIÓN SEMIAUTOMATICA DE TESAUROS

Juan Lloréns Morillo [llorens@gti.uc3m.es]
Manuel Velasco [manu@gti.uc3m.es]
Jorge Morato [jorge@gti.uc3m.es]
GTI (Grupo Tecnologías de la Información). Facultad de Informática. Univ.Carlos III (Leganés, Madrid)
José A. Moreiro
Facultad de Biblioteconomía y Documentación. Universidad Carlos III (Getafe, Madrid)

RESUMEN

El objetivo del GTI es la generación semiautomática de tesauros mediante el análisis de un corpus. Tras ensayar distintos métodos de clasificación de la información, desde co-ocurrencia de términos a redes neuronales, se mostró necesaria la creación de nuevos indicadores que aportasen información adicional a la ya suministrada por el tesauro. La presentación de estos indicadores, y su previsible potencial, es la meta de la presente comunicación. El objetivo es reaprovechar el gran volumen de datos necesarios para realizar la clasificación y emplearlos en dos campos distintos: por un lado la validación del tesauro y por otro la creación de indicadores que nos indiquen a-priori la creatividad del texto dentro de nuestro corpus. La estructuración y etiquetado previo del texto parecen en estas circunstancias un paso necesario para poder estudiar posteriormente el resultado del conjunto de parámetros medidos en el set de documentos. La novedad se estudia desde un enfoque multidimensional: análisis lingüístico y del formato de los textos, estudio del tesauro generado, y la creación de indicadores ad-hoc. Al tiempo, se miden distintos parámetros en el tesauro para validar el tesauro autogenerado. Para el análisis matemático de los datos, se usan análisis multivariante y de las componentes principales. Una evaluación del programa esta actualmente en curso.

INTRODUCCIÓN

Durante los últimos años, se ha incrementado la necesidad de investigar sobre herramientas que mejoren la recuperación de documentos en formato electrónico. La utilización de un tesauro que posibilite la recuperación en lenguaje controlado se ha ido abriendo paso como una de las opciones más prometedoras. Aunque, esta aparente ventaja no lo es tanto cuando vemos a los tesauros como un sistema demasiado rígido en el transcurso del tiempo. Los nuevos términos y nuevas relaciones no son siempre ubicables en el tesauro pues los marcos científicos tienden a quedar obsoletos en plazos de tiempo relativamente cortos. Lamentablemente, la elaboración manual de tesauros es una tarea no solamente costosa, sino lenta.

Desde el 1992, el objetivo del Grupo de Tecnologías de la Información (GTI) ha sido la generación de tesauros de forma (semi)automática. Aprovechando un entorno de generación automática de tesauros, se ha encontrado interesante no sólo localizar y determinar nuevos conceptos, sino reestructurar el tesauro para acomodarlo a los nuevos cambios que precisa el corpus. En este documento se van a poner de relieve los indicadores que se muestran como más prometedores para localizar esta información en los textos.

Los tesauros representan un conjunto de atributos y relaciones entre conceptos relativos a un particular campo de conocimiento, que puede ser aplicable a los entornos más variados.

De igual forma, un tesauro se puede definir como un área de conocimiento, actividad, interés o aplicación, que hace referencia a una doble realidad, por un lado, el tesauro como referencia al concepto en sí mismo y, por otro, el análisis que representa el marco que posibilita la clasificación y la recuperación (Velasco, 1997). Algunos autores (Arango, 1991), (Prieto-Díaz, 1991), (Llorens, JA1236)

presentan técnicas que posibilitan el desarrollo de un análisis del tesauro (DRACO, DARE, FODA...), y que en la caso de Velasco et al., incluso se presta a la automatización.

La estructura de representación de tesauros (ERD), con la que operaremos, fue creada según el protocolo de la ISO (ISO2788) (Aitchinson, 1987), (Comm, 1993) para tesauros monolingües, en el que se proponen herramientas para la construcción semiautomática de un tesauro estructurado para un conjunto de documentos. El tesauro incluye también la clasificación decimal (Dewey, 1979) preparada para una clasificación facetada (Ranganathan, 1967) dentro de una estructura llamada Árbol de Areas Temáticas (AAT) (Llorens, 1996).

Tradicionalmente, ha sido la Infometría la encargada de identificar y medir las diferentes agrupaciones dentro del set documental, ya sea por análisis de referencias, citas o coaparición de palabras (Callon, 95). Aunque estos indicadores cienciométricos tienen una amplia difusión, la literatura sobre indicadores cualitativos continua siendo escasa (Rip, 1997). Una de las principales características que, definen la calidad del documento es la novedad (Buchholz, 1995). En las próximas páginas, se describe un sistema que, complementando al ya existente de clasificación, facilita información sobre la creatividad de los documentos (Egghe, 1990). Para poder encarar este problema, se hacen necesarias herramientas para *medir la calidad* documental, aunque, auxiliados por otras técnicas info-bibliométricas que nos ayuden en esta tarea (Polanco, 1995).

A continuación, se describe el proceso con el que se trabaja actualmente. En primer lugar, se detallan las manipulaciones previas del set documental. Posteriormente, se muestran las herramientas utilizadas para validar el tesauro y estimar la creatividad, para, por último, realizar el análisis de datos.

Aun excediendo los objetivos de la presente comunicación, se ha considerado de interés hacer una breve referencia a los algoritmos de clasificación y al de *stemmer*, ya que la elección de uno u otro método condiciona los resultados posteriores.

1. MANIPULACIONES PREVIAS AL CÁLCULO DE VARIABLES

1.1. SELECCIÓN DOCUMENTAL

La selección documental es un paso importante para conseguir una buena representación del tesauro. Antes de realizarla es necesaria una cuidada planificación, obteniendo así resultados más fiables en cálculos posteriores. El número de documentos que forman el corpus debe ser calculado, ya que un número alto de documentos no siempre, lleva aparejada una mejora en la representación del corpus, pues las redundancias y el ruido aumentan proporcionalmente al incrementarse el número (Van Slype, 1987). El corpus deberá estar compuesto por material textual preferentemente en formato electrónico: artículos, libros, manuales, bases de datos documentales, etc. que deben ser seleccionados tanto por el documentalista como por el informático (Prieto-Díaz, 1990).

En principio, las referencias y los artículos procedentes de revistas científicas presentan ventajas debido a dos factores, por un lado, el lenguaje empleado y, por otro, el poseer una estructura bastante uniforme. El lenguaje científico se caracteriza por poseer un vocabulario objetivo, normalizado (se evita la sinonimia), universal y con una previsible calidad (*peer-review*). En nuestro caso, se han seleccionado artículos médicos a texto completo, de la base de datos med-line. La búsqueda se realizó con un conjunto discreto de palabras-clave, dentro de las suministradas por el tesauro de med-line, Medical Subject Headings (Medical, 1998).

El corpus debe ser lo más homogéneo posible y es preferible que estén escritos en un área, época, idioma y disciplina lo más similar posible para obviar, en lo posible, efectos colaterales no deseados (Cleveland, 1990). Para ello, la búsqueda se limitó a dos revistas de los años 96 y 97, de habla inglesa. En cualquier caso, el número de autores no debe de ser demasiado alto.

Existen varios motivos por los que se prefiere el formato electrónico, así entre otros, están:

- Utilizar el incremento de material en formato electrónico.
- Aprovechar los estándares para etiquetado (por ejemplo, SGML o HTML) (Polanco, 1995).
- Evitar problemas con la digitalización.

1.2. TRATAMIENTO PREVIO DE LOS DOCUMENTOS

Se ha desarrollado un componente DCOM (Brookschmidt, 1995), para acceder a la información a través de múltiples interfaces. Cada interface puede recuperar de los documentos un conjunto de datos diferente, mediante su relación con tecnologías del tipo OLE Automation (OLE, 1996) e iFilter API, entre otras, pudiendo acceder a archivos del tipo de .txt, .doc, .html, .xls, .ppt, .pdf....

En una primera etapa, se realiza una normalización del documento para que guarde la máxima homogeneidad posible con el resto del corpus. En el caso concreto de Microsoft Word, se extrajeron los datos bibliográficos de las propiedades del archivo. A todos los títulos y subtítulos de capítulo, se les ha asignado un determinado estilo de texto. No ha sido posible automatizar en todos los casos la extracción del documento de algunos de estos elementos por lo que, actualmente, se esta trabajando en introducir manualmente estos datos, por ejemplo, factor de impacto de la revista, nombre de la revista, afiliación de los autores o tipo de documento. Para la identificación del idioma del documento, se utilizó la herramienta de definir idioma.

1.3. EXTRACCIÓN Y ALMACENAMIENTO DE DATOS

Un área de creciente estudio es la exploración del contexto con el fin de extraer información. El objetivo es capturar información lingüística y del formato del texto, para poder estructurar un documento (Lazariņis, 1998), dividiéndolo en distintos apartados como conclusiones, métodos,.... Como se ha comprobado en observaciones cognitivas los indizadores profesionales utilizan marcadores textuales, estructurales y semánticos (Smith, 1994). Como también, tiene base cognitiva el considerar, que si un autor quiere recalcar determinada parte del documento empleara diferentes formatos de texto (Berri, 1996). La forma en que se realiza este proceso es relativamente simple; por un lado, se localiza en el texto un conjunto discreto, y previamente tabulado, de términos del tipo: conclusión, resultado, método, resumiendo, concluir, etc.; por otro lado, se le da diferente peso según su estilo y su formato (título, subtítulo, negrilla, etc.). Unas pocas reglas sobre la ubicación complementan el proceso, como el que la probabilidad de que se trate de las conclusiones, aumenta según nos aproximamos a la parte final del documento. El fin es localizar e identificar esta información para etiquetarla en el texto. Las posibilidades que abre esta opción son, entre otras, la de poder comparar independientemente los descriptores de las distintas secciones de los documentos.

Si se observan los estudios sobre bibliometría, se puede comprobar que a menudo se circunscriben a un mero análisis de agregados de las referencias del corpus, casi siempre limitado a las revistas del Science Citation Index. A causa de esta evidente limitación, en los últimos años, estos análisis están siendo abandonados en favor de los análisis de co-ocurrencia de términos, lo cual nos provee de un nuevo enfoque para analizar los textos. En el presente proyecto, se ha estimado que, lejos de sustituir al método clásico, el método de coaparición de términos, lo complementa. A tal fin, se han programado reglas para manejar las referencias de la bibliografía. De este modo, la fuente, autoría, coautoría, cooperación, idioma y fecha, son extraídas automáticamente de cada una para su posterior análisis.

También, se han necesitado un conjunto relativamente alto de reglas para manejar la información especial, como: siglas, citas a referencias bibliográficas, tablas, información extraída de los pies de las figuras, fórmulas, fechas, información gráfica, referencias, tipos de términos, listas e hipervínculos.

1.3.1. ASIGNACIÓN DE PESOS POR LOCALIZACIÓN Y FORMATO

En el análisis de co-ocurrencias, la elección de porciones mayores de texto abre grandes expectativas a la hora de asignar pesos. Para ello, hemos definido como frase al conjunto de palabras, que no siendo fórmulas, tablas, direcciones o acrónimos, esta limitado por tabuladores, principios de párrafo o puntos. En este sentido, los párrafos están confinados por retornos de carro. Una de las ventajas de utilizar artículos de revistas científicas, es que están a menudo estructurados en capítulos, pudiendo asignar distintos pesos, a los descriptores según procedan de un lugar u otro del documento.

Aunque con frecuencia sólo se indizan las palabras contenidas en el resumen y el título, hemos considerado interesante incluir otros aspectos. En lo concerniente a las palabras se tiene tanto en cuenta el formato del texto, como los diferentes estilos en los que esta definido el documento.

Por otro lado, se ha creído de interés ver el uso más o menos alto de pronombres en el texto. El objetivo es utilizarlo en la validación del tesauro, ya que los pesos de los descriptores se establecen según el número de apariciones en el texto. Por lo tanto, los pronombres que sustituyen a los verdaderos descriptores, al igual que las negaciones, provocan una disminución en la calidad del tesauro generado. A tal fin, se han creado dos tablas en SQL con negaciones y pronombres tabulados para su localización en el texto.

Las citas dentro del texto a las referencias bibliográficas merecen una mención aparte. El peso de cada referencia en el corpus ha sido ponderado según el número de veces que ha sido citada. Tampoco, hay que olvidar que la existencia, en una misma frase, de una partícula negativa en combinación con una referencia, se ha considerado tradicionalmente como una desventaja en la fiabilidad de los análisis de referencias. En nuestro caso, queremos saber si efectivamente el peso de la referencia debería estar ponderado por este factor.

1.4. WORDER

Uno de los problemas que tiene que encarar cualquier programa de indización es el de las variantes morfológicas. El no reducir cada una de las variantes a una raíz común da pesos inferiores a los deseados para ese término (Hull, 1996). Para encarar este problema se creó el módulo de stemmer (Worder).

Las variantes morfológicas procedentes de las reglas de derivación y flexión de términos, se han tratado con el módulo, obteniendo formas normalizadas. Por ejemplo, programar, programación y programas se normalizan a programa.

Este cambio se realiza seleccionando en el término no normalizado la terminación y sustituyéndola por una forma normalizada. Para determinar el tamaño de esta terminación, se extraen los caracteres hasta el primer carácter común entre el término normalizado y el no normalizado. Aunque la extracción de los prefijos no ha dado, por el momento, los resultados esperados, también se está programando con un propósito meramente experimental.

Cada sustituto incluye información sobre el fin que permite ese cambio, por ejemplo, adjetivación, substantivación, etc. La selección del tipo de regla es optativa, permitiendo un control sobre la normalización. Cada palabra así obtenida es comparada con las existentes en el tesauro. La ventaja, frente al método tradicional de buscar la presencia en un diccionario, es que el tesauro suele estar más contextualizado que los diccionarios.

1.5. BASES DE DATOS

1.5.1. BASES DE DATOS CREADAS PREVIAMENTE

Con anterioridad a la extracción de datos, determinadas partículas se han tabulado mediante tablas SQL. Estas bases de datos se van a confrontar con los elementos extraídos de los documentos, para ver así su frecuencia relativa en los textos. Así, se crean las siguientes tablas:

- Pronombres.
- Negaciones.
- Palabras vacías, son términos con poco valor informativo, pero con alta frecuencia de aparición.
- Términos asociados al concepto de novedad (p.e. nuevo, creación, ...).
- Términos para la identificación del capítulo, dentro de la estructura de cada documento (p.e. introducción, método, resumen, concluyendo, ...).

1.5.2. BASES DE DATOS CREADAS TRAS LA EXTRACCIÓN DE DATOS

Varias bases de datos se van a crear tras el proceso de extracción y almacenamiento de datos. Será principalmente en estas tablas sobre las que se realice posteriormente la estimación de pesos por el programa estadístico. De esta manera, se generan las siguientes bases de datos:

- Términos: contiene información sobre la frecuencia de cada término, posición, formato, estilo, número de frase, párrafo, capítulo y documento en el que aparece. El campo en el que se encuentra, por ejemplo, tabla, pie de figura, etc. también se almacena.

- Referencias: autores de las referencias, revistas, número de referencias en un idioma diferente al del documento en que aparece, número de referencias, número de veces que se cita la referencia en el texto y número de veces que esa cita aparece junto a una negación.
- Set de documentos: la información bibliográfica de cada documento es extraída de su ficha y tabulada, en los siguientes campos: nombre de los autores, número de autores por documento, instituciones, número de instituciones y tipo de instituciones (pública, privada, universitaria o foránea).
- Descriptores del tesauro generado. Se almacena el lugar donde aparecen (capítulo, campo, etc.) y su formato. Cuando aparece con otro descriptor del documento en la misma frase, se acumula también la combinación.
- Inclasificados, listado de términos que sin ser palabras vacías y sin llegar a una frecuencia suficiente para ser descriptores, aparecen con cierta frecuencia en la colección.

2. CREACIÓN DEL TESAURO

2.1. HERRAMIENTAS PARA GENERAR EL TESAURO

En una anterior etapa, se realizaron filtrados como el IDF (Inverse Document Frecuency) (Cleveland, 1990), basada en la ley de Zipf (Zipf, 1972), y el método de n-grams (Cohen, 1995). Por otro lado, las palabras vacías se suprimieron (Frakes, 1992). Para determinar los distintos grados de equivalencia y jerarquía, se utilizó una red neuronal (Llorens, JA1236).

2.2. VALIDACIÓN DEL TESAURO

2.2.1. VARIABLES PARA VALIDAR EL TESAURO

A continuación, se define un conjunto de variables con las que se pretende evaluar su impacto sobre la validación del tesauro y sobre el apartado de novedad:

Especificidad:

Es el número de términos en el tesauro más específicos que los descriptores del artículo, dividido entre el número de descriptores. A parte de una mala representación, un valor alto podría deberse o a una mala adecuación a ese tema por el tesauro, o a un documento generalista.

Coeficiente de dispersión (n° temas) (Cd):

Es el número de ramas o subárboles que tiene un documento sobre el total de ramas de todo el tesauro en el nivel f(x). Y se expresa como el sumatorio del número de nodos de nivel n, de los que depende al menos un descriptor del documento, dividido entre el número total de nodos en el tesauro a ese nivel.

Coeficiente de clasificación (Cm):

Es el número de nexos que separan a cada descriptor del documento de cada uno de los demás descriptores del documento.

Coeficiente de Llorens (Hl):

Se calcula considerando la posición jerárquica de los descriptores del documento respecto a los nuevos descriptores creados.

$C_i = \Sigma C_m - 2(N_d^2 - 2N_d + 1)$, donde:

Cm es el sumatorio de los coeficientes de clasificación de un documento.

C_i es 0 cuando todos los descriptores de la jerarquía se pueden encontrar en el documento y si la jerarquía tiene sólo 2 niveles diferentes con solo un descriptor raíz.

N_d número total de descriptores en el documento

H_l es el Ci dividido entre f(x); donde f(x) es $x^2 / x+1$ y donde x es el número de niveles de jerarquía

Este coeficiente se puede utilizar para medir la homogeneidad del tesauro, es decir la relación entre el número de descriptores y el número de niveles de jerarquía.

La utilidad de estas ratios tiene un doble objetivo, por un lado, valores elevados de estos coeficientes, nos pueden indicar una representación del tesauro poco adecuada. Por otro lado, ante nueva información, es previsible que también aumenten los valores de los coeficientes, debido a cambios en los descriptores y enlaces.

3. MEDIDA DE LA NOVEDAD

Se han diseñado varios métodos para analizarla:

3.1. Análisis de contenido

Se ha utilizado el análisis de contenido (Rosenthal, 1994), como un método para distinguir los documentos más creativos de los menos creativos (Amoroso, 1995). Como en el apartado de *Extracción y Almacenamiento de Datos*, en el que se aprovechaba la presencia de determinados términos para estructurar el documento, se puede medir la novedad. El

método ha utilizado la comparación de las palabras del texto con una lista de palabras significativas en el apartado de novedad. Algunas de estas palabras son: creación, diseño, generación, idea, innova, original, producto, ... A partir de determinado nivel de coincidencias entre la lista y las palabras de un documento, se decide su posible importancia en relación a este apartado. Con un método similar, Haas (Haas, 1995) realizó con éxito una identificación de artículos experimentales.

3.2. Análisis del tesauro generado

Desde el punto de vista tradicional, los tesauros se definen como un sistema rígido, en que nuevos conceptos y relaciones suelen ser un gran obstáculo. La generación automática de tesauros supone una considerable ventaja, ya que el problema que suponen los nuevos términos es abordable más fácilmente. Esto nos permite tratar a los tesauros como un sistema dinámico en el tiempo.

Cuando una vez generado el tesauro, se vuelve a analizar un nuevo corpus sobre el mismo tema pero de distinta época, se pueden dar varios casos. Si al comparar los dos tesauros se ha generado un nuevo agregado que se considera significativo, puede ocurrir que los descriptores de este nuevo agregado estuvieran ya presentes o no. Si no estaban presentes, se deberán reubicar en el antiguo tesauro (Chen, 1995). El número de estos descriptores *inclasificados* nos da una importante información referente a la novedad. Otra posibilidad que se puede dar es que los agregados sean practicamente idénticos pero la *intensidad* de los enlaces entre estos haya cambiado. En este caso, el análisis estadístico sobre el tesauro de las ratios, que se ha descrito en el apartado de *Validación del Tesauro*, tiene un indudable valor para detectar documentos novedosos.

3.3. Co-ocurrencia de descriptores dentro de una misma frase

La utilización de frases para la indización ha tenido un gran desarrollo durante los últimos años, gracias al avance sobre procesamiento del lenguaje natural. En este contexto, la utilización de un método similar al de concurrencia de términos ha sido propuesta (Callon, 1995). La idea que se propone consiste en determinar conjuntos de descriptores precoordinados, mediante su concurrencia reiterada en una misma frase.

Así se pretende analizar las variaciones de estos descriptores múltiples, ya que cuando un área tiene un auge creciente, surge un número creciente de variantes. La co-aparición con otros términos, en la misma frase, forma unidades cada vez mayores. Estas difieren del original en variaciones debidas a coordinación, permutación o inserción de los nuevos términos entre los elementos del antiguo descriptor múltiple (Muller, 1997). No sólo es importante saber *cuántas* variantes se han generado, sino también *cuándo*.

3.4. Indicadores infométricos

Se han definido varios parámetros ha calcular en el corpus, en cada documento y en cada una de las secciones en las que hemos estructurado el documento (Velasco, 1997).

3.4.1. Sobre el texto

- *Rareza*: Porcentaje de términos que aparecen menos de n veces.
- *Frecuencia y número de descriptores, por capítulo.*
- *Frecuencia y número de citas* a las referencias.
- *Frecuencia, de palabras inclasificadas*, con relación al número total de palabras.

3.4.2. Sobre propiedades docuemento

- *Coautoría* (número de autores por artículo) y colaboraciones (instituciones por artículo).

3.4.3. Sobre referencias bibliográficas

- *Número de referencias y su frecuencia.*
- *Vida media* de las referencias.
- *Frecuencia de autocitas* respecto al número total de referencias. Autocita, es el número de referencias en que su autor coincide con el autor del artículo.
- *Capacidad idiomática* (número de referencias en un idioma por cien entre el número total de referencias). Para determinar un índice de aislamiento de las referencias bibliográficas se ha calculado el porcentaje de referencias en un idioma diferente al del cuerpo del artículo. Para ello, se ha estimado la coincidencia entre las palabras de la referencia y las de un diccionario electrónico del idioma en el que está escrito el artículo. El número de palabras a identificar se calcula como el

- 60% de las que componen la referencia bibliográfica.
- *Frecuencia de aparición de cada revista.*
- *Peso de cada referencia* (número de referencias de un artículo que están repetidos más de n veces en el texto del documento por cien, dividido entre el número de referencias del documento)
- *Coaparición de referencias* (número de documentos en que aparecen a la vez la referencia i y j)[2/] (número de documentos con la referencia i * número documentos con la referencia j))

3.4.4. Sobre agregados

- Valor de *Centralidad* y de la *Densidad* (para definiciones, ver Callon, 1995)

3.4.5. Temporales

- *Índice de transformación* entre el mismo agregado en dos momentos diferentes (número de descriptores en común entre los agregados i y j al cuadrado entre el número total de descriptores en el agregado i por el número de descriptores en el agregado j)
- *Cambio en inclasificados*: $A=a_2-a_1$, donde: a es el número de palabras inclasificadas dividido entre el número de palabras inclasificadas más el número de palabras clasificadas (siendo a_1 y a_2 son valores de a en dos momentos distintos)

4. ANÁLISIS DE LOS DATOS

Un análisis de regresión, con las diferentes combinaciones de variables, es el método estadístico empleado para analizar los datos anteriores. Se aplica, tanto para validar el tesauro, como para ver la novedad de los nuevos documentos, dando así pesos (Everitt, 1991).

En el análisis de regresión, representado por: $Y=X\beta+u$ (Chatterjee, 1991), donde Y es la variable dependiente, X es la matriz formada por el valor de las variables en cada observación y β son los parámetros estimados por los coeficientes de regresión b_0, b_1,....Donde $b=(X'X)^{-1}X'Y$.

Actualmente, se está evaluando la importancia de los valores de x para predecir los valores de y, y así analizar los efectos que tiene sobre la variación de los valores y. El programa elegido ha sido el SPSS, debido a su compatibilidad con el resto de las aplicaciones.

5. CONCLUSIONES

La literatura sobre creatividad es actualmente escasa, ya que presenta evidentes dificultades a la hora de asignar un valor objetivo. Aunque la evaluación estadística del proyecto está actualmente en curso, es previsible una gran cantidad de ruido debido al hecho multidimensional que se pretende medir.

La utilización de varios indicadores que muestren los distintos aspectos del concepto de creatividad, ofrece, en la práctica, una mejor representación de la realidad, al tiempo que, disminuye la posibilidad de manipulación por parte de los autores.

En lo relativo al análisis estadístico, se han realizado frecuentes críticas por la aplicación sistemática de este tipo de estadísticas en el campo de las ciencias sociales (Haitun, 1982), debido a la ausencia de normalidad de la mayoría de las distribuciones analizadas. Aunque, hoy por hoy parece ser la única solución factible (Ferreiro, 1993).

En una anterior etapa, los resultados con los algoritmos de clasificación ofrecieron resultados esperanzadores (Llorens, JA1236). La necesidad de un cuidadoso análisis de los resultados con una muestra de mayor volumen permanece pendiente. Así, como su ajuste con las ratios definidas más arriba.

REFERENCIAS

Aitchison, J. (1987) "Thesaurus Construction: A Practical Manual". ASLIB.

Amoroso, Donald L. and Couger J. Daniel. (1995). "Developing information systems with creativity techniques: an exploratory study". Proceedings 28th Ann. Hawaii Intern. Conf. on Sys. Sc. 720-728

Arango, G. & Prieto-Díaz, R. (1991) "Domain Analysis and Software Systems Modeling". IEEE Computer Society Press.

Berri, J. et al.. "A Linguistic Method For Text Filtering". (1996). Procesamiento del Lenguaje Natural 19: 159-165.

Buchholz, K. (1995) "Criteria for the analysis of scientific quality". Scientometrics 32(2): 195-218.

Callon, Michel, Jean-Pierre Courtial y Hervé Penan. (1995). "Cienciometría: la medición de la actividad científica: de la bibliometría a la vigilancia tecnológica" Gijón : Trea. 110 p.

Chen, H., Yim, T., Fye, D. & Schatz, B. (1995) "Automatic Thesaurus Generation for

an Electronic Community System". Journal of the American Society for Information Science 46 (3): 348-369.

Cleveland, D. B. & Cleveland, A. D. (1990) "Introduction to Indexing and Abstracting". Libraries Unlimited.

Cohen, J. (1995) "Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting". Journal of the Amer. Society for Information Science. 46(3).

Commission of the European Communities (1993) "Thesaurus Guide: Analytical Directory of Selected Vocabularies for Information Retrieval". Luxembourg: Office for Official Publications of the European Communities.

Dewey, M. (1979) "Decimal Classification and Relative Index" Forest Press Inc.

Egghe, Leo & Rousseau, Ronald. Introduction to informetrics: Quantitative methods in library documentation and information science. Amsterdam: Elselvier Science Publishers, 1990. 447 p.

Everitt, Brian S. And Dunn, Graham. Applied Multivariate Data Analysis. London [etc.] : Edward Arnold, 1991. 304 p.

Ferreiro Aláez, Luis. Bibliometría : (análisis bivariante). Madrid: EYPASA, 1993. 480 p.

Frakes, W. B. & Baeza-Yates, R. (1992) "Information Retrieval: Data Structures & Algorithms". Prentice Hall.

Haas, Stephanie W., Sugarman, Jeremy and Tibbo, Helen R. (1996). A text Filter for the Automatic Identification of Empirical Articles. Jasis 47 (2): 167-169

Haitun, S.D. (1982). Stationary scientometric distributions, Scientometrics, 4; Part I, 5-25, Part II, 89-104, Part III, 181-194

Hull, David A. (1996) "Stemming Algorithms: A case study for detailed evaluation". JASIS 47(1): 70-84.

ISO 2788 -1986 (E). "Guidelines for the establishment and development of monolingual thesauri". 2nd ed. (1986)-11-15 UDC 025.48. ISO 2788.

Lazarini, Fotis. Combining Information Retrieval with Information Extraction for Efficient Retrieval of Calls for Papers. In: 20th Colloquium of the British Computer Society's IRGS (Grenoble, 1998), Draft proceedings. Mark D Dunlop, 1998: 162-174.

Lloréns, J. (1996): "Definición de una Metodologia y una Estructura de Repositorio orientadas a la Reutilización: el Tesauro de Software". Ph. D. Univ. Carlos III de Madrid.

Llorens, J., Moreiro, J. A., Amescua, A., Martinez, V. & Velasco, M. "Automatic Generation of Domain Representations using Thesaurus Structures". JASIS (admitido para publicar, número de referencia JA1236).

Medical Subject Headings (MeSH) (1998). http://www.nlm.nih.gov/mesh/mtrees.html

Muller, C. et al. "Acquisition et structuration des connaissances en corpus: éléments méthodologiques" [Knowledge Acquisition and Structuration from Corpora]. Nancy: INRIA research papers, 1997. 45 p.

Polanco, X., Grivel, L. & Royauté, J. How to do Things with Terms in Infometrics: Terminological Variation and Stabilisation as Science Watch Indicators. In: Proceedings Fifth International Conference on Scientometrics and Informetrics. Medford (NJ): Learned Information, 1995: 435-444.

Prieto-Díaz, R. (1990) "Domain Analysis: An Introduction". ACM Sigsoft. Software Engineering Notes. Vol 15(2).

Prieto-Díaz, R. (1991) "Implementing Faceted Classification for Software Reuse". Communications of the ACM. Vol 34(5).

Ranganathan, S. R.(1967) "Prolegomena to Library Classification" Asian Publishing House. India.

Rip, A. (1997) "Qualitative conditions of scientometrics: the new challenges". Scientometrics 38(1): 7-26

Rosenthal, Robert and Rosnow, Ralph L. Essential of behavioural research: methods and data analysis. In: McGraw-Hill Ser. in psychology. N.Y.: McGraw-Hill, 1984. 519 p.

Smith, Philip, J. et al. Computerised tools to support document analysis. In: Challenges in indexing electronic text and images. Medford (NJ): ASIS, 1994

Van Slype, G. (1987) "Les Langages d'indexation: conception, construction et utilisation dans les systèmes documentaires". Paris Les Editions d'organisation.

Velasco, M., Lloréns, J. & Martínez Orga, V. (1997) "Generación automática de representaciones de tesauros". II Jorn. Ingen. Software. JIS97. San Sebastián. Spain.

Zipf, G. K.(1972) "Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology". Haffner. New York.