Reconocimiento de títulos de artículos no concordantes con el contenido a través de la utilización de palabras clave

Manoel Palhares Moreira
Sergio Murilo Stempliuc
Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte (Brasil)

Resumen

Se aborda la detección de títulos que no representan suficientemente bien el contenido de los artículos científicos a traves de la comparación de los mismos con las palabras clave seleccionadas por el autor o el editor. Se partió de la hipótesis de que artículos con al menos una de sus palabras clave en el título poseen contenido concordante con el mismo. Se escogieron 300 artículos de dos revistas brasileras disponibles en la Web. En primer lugar, y mediante ordenador, se verificó la recurrencia de las palabras clave indicadas por los autores en los títulos de sus artículos: en el 23% de los casos (69 artículos) ninguna palabra clave fue encontrada en el título. La verificación de la recurrencia de estas palabras clave en las demás unidades de estos textos mostró que el 13% (9 textos) no poseían ninguna de sus palabras clave en el cuerpo del texto; en el 36% (25 textos) no fueron encontradas en los resúmenes; y, en el 58% (40 textos), en las referencias bibliográficas. En una segunda etapa se realizó un análisis cualitativo de los datos. Se verificó que en algunos de estos casos el problema era la designación de las palabras clave, por lo que se compararon las del título con las del texto completo. Solo un 1% de los artículos planteaban problemas debido a la perspectiva creativa de los autores.

Palabras clave: Organización de la información. Títulos de artículos científicos. Contenido de artículos científicos. Tratamiento de la información. Recuperación de información.

Abstract

A methodology for detecting titles that are not reliable surrogates of the content of scientific articles is proposed by comparing them with the keywords or descriptors of the article. It is based on the hypotheses that papers with at least one of their keywords included in the title have a corresponding content. 300 articles from two Brazilian web-magazines were used. First, the occurrence of keywords was

checked: in 23% of the cases (69 papers) no keyword was found in the title. Such checking in the other texts showed that 13% (9 texts) had none of the keywords in the text; in 36% (25 texts) they were not found in the abstracts; and, in 58% (40 texts), in the bibliographical references. Next, a qualitative analysis of data was carried out. It was verified that in some cases the problem was keywords indication. As a proposed solution, a checking routine was done, for the frequency of all words used in the papers but stop-words. After that, came the checking of the word variation in the indicated keywords. The conclusion was that keywords in paper titles is a factor that may be observed during the submission process for the checking of title with paper content, and that digital magazines might control vocabulary built form such words. That measure would make it easier for authors in the title creation and keyword indication. Such procedures would facilitate search and finding for documents before users' needs.

Keywords: Information organization. Controlled vocabulary. Title of scientific papers. Content of scientific papers. Information treatment. Information retrieval. Keywords.

1. Introducción

El período entre finales de la década de los cuarenta y el inicio de los años cincuenta estuvo marcado por el aumento de las formas de producción y difusión de la información y por el deseo de obtener esa información. Los años de guerra impulsaron el trabajo y el desarrollo en diversas áreas; todo esto hizo que las personas estuvieran constantemente preocupadas por la defensa de su país.

El desarrollo tecnológico tuvo una importancia decisiva en la resolución de la segunda guerra mundial y afectó sustancialmente a la orientación dada a la industria, lo que llevó a que las directrices políticas de varios países contemplaran el tema del desarrollo de la ciencia y la tecnología. En consecuencia, creció la producción científica y se hizo evidente que era necesario acelerar el proceso de difusión del conocimiento generado.

En 1945, Vannevar Bush publicó un artículo en el que señalaba la necesidad de hacer accesible el conocimiento generado en ciencia y tecnología, sugiriendo la aplicación de tecnología de la información como una solución posible para la recuperación de información frente al acelerado crecimiento de la producción científica (Saracevic, 1995; Busch, 1945). Muchos otros investigadores compartieron esa preocupación y se entusiasmaron con la posibilidad de soluciones tecnológicas para el problema.

En los años cincuenta, la recuperación de información gana fuerza entre la comunidad científica y, en algunos países, como Estados Unidos, donde la solución para este problema pasa a tener el apoyo del Gobierno, se dirige en un primer mo-

mento a la explosión informativa en ciencia y tecnología, para luego trasladarse a otras áreas del conocimiento humano.

En Brasil también hubo un salto en el desarrollo institucional de la ciencia. Con el ejemplo del modelo y del consenso norteamericano, surgieron en las décadas de los cincuenta y los sesenta dos agencias de fomento a la investigación en el país: el Conselho Nacional de Pesquisa (CNPq), creado en 1951, y la Fundação de Amparo á Pesquisa do Estado de São Paulo (Fapesp), instituida en 1962.

La segunda guerra mundial marcó a la humanidad, influenció el rumbo y las opciones de los países y trajo consecuencias para la vida de todos. Esta herencia del período de guerra hizo crecer la producción científica y representó para las bibliotecas un desafío en relación con los procesos de indización y recuperación de información. Hasta ese entonces era posible organizar los documentos del acervo, constituido en libros casi en su mayoría, a través de instrumentos previamente construidos. El bibliotecario era el responsable de la custodia de estos bienes documentales, aunque pretendiese ser un intermediario entre la necesidad de información y el acervo disponible. Surgieron también otros géneros y soportes documentales, principalmente artículos de periódicos científicos e informes de investigaciones, lo que representó el comienzo de la era de la especialización del conocimiento (Foskett, 1973; Dodebei, 2002).

En los últimos años, el avance de la tecnología posibilitó una mayor agilidad en el acceso y la difusión de la información en los diversos niveles de la sociedad. La conjunción de las conquistas de la computación y de las telecomunicaciones hizo posible lo que se denomina *tecnología de la información* (Oliveira y Noronha, 2005), que en el contexto de los documentos y producciones científicas inauguró, con formato digital, un nuevo soporte para el registro de información.

También como fruto de esa tecnología, en los años noventa surgió la World Wide Web (Web), que iba a convertirse en el gran faro mundial de las informaciones.

La posibilidad del soporte digital, junto con la tecnología web, permitió que se introdujeran nuevas formas de difusión y recuperación de información mediante la disponibilidad de los documentos en red y el desarrollo de sistemas de recuperación de información, haciendo posible la llamada *sociedad en red* (Castells, 1999), donde la información es producida y almacenada en un lugar diferente de donde el usuario la recupera.

La desmaterialización de los objetos informacionales abrió nuevas perspectivas a la representación y a la recuperación de los documentos. Esas tecnologías alteraron también el modo de comunicación entre las personas y aportaron modificaciones a la comunicación científica; sin embrago, sin dejar a un lado el camino ya recorrido, deben al mismo tiempo repetir fórmulas ya utilizadas en la recuperación de información.

Si hoy tenemos fácil acceso al conocimiento, alabado sea el lenguaje, pues a través de él, principalmente, nos comunicamos con el grupo en el que estamos inmersos, con el que trabajamos o estudiamos. El lenguaje es la expresión de nuestra lengua natural o nativa, nuestro modo de hablar normalmente, el que hace comprensible nuestra comunicación y nos saca de nuestros propios límites.

El lenguaje también es el medio por el cual el hombre crea y amplía su conciencia, pues hace posibles actos como simbolizar, conceptualizar o clasificar (Murriel, 1998). En él nos apoyamos para anclar nuestras culturas, nos identificamos como pueblo de un lugar y una época determinados. Este es el poder de la lengua, manifestado y modificado en lo cotidiano, en nuestro modo de vivir el *cultus* y fortalecer nuestra propia cultura.

En la Web, los sistemas de recuperación de información, en concreto las máquinas de búsqueda, se valen de un lenguaje natural para los procesos de recuperación de información. El lenguaje natural es el que una comunidad utiliza cotidianamente en su comunicación escrita o hablada. Y es también el utilizado en los textos científicos.

Se consideran lenguajes *de indización* los instrumentos de representación de la información para la indización o el almacenamiento y la recuperación de documentos.

Los periódicos científicos nacieron del intento de proporcionar una mayor comunicación entre la comunidad científica. Aunque estén elaborados en un lenguaje natural, obedecen a un determinado formato, con normas para su construcción. Y algunas partes específicas de un texto son prioritarias tanto para quien las escribe como para quien las lee.

El presente trabajo presenta experimentos realizados con artículos de dos periódicos brasileros. Los títulos se convirtieron en el objeto de estudio y fueron comparados con el contenido de los artículos. El esquema del trabajo es este: primero se presenta el problema, los objetivos de la investigación y la metodología empleada; en el punto siguiente se recupera la trayectoria histórica de los periódicos científicos y las normas generales que componen su estructuración; luego se presenta el instrumento desarrollado para el experimento y el análisis de los datos obtenidos. Y, a modo de conclusión, se expone la lectura de este recorrido y sus expectativas de continuidad.

2. El problema investigado

Durante años se llevaron a cabo múltiples investigaciones sobre recuperación de información. Las ideas y experiencias de los años cincuenta y sesenta se transformaron en bases de datos propios para recuperación de información y en servicios y sistemas especializados en el tema. En los años siguientes, el advenimien-

to de la tecnología de redes cambió la trayectoria, pero reforzó estas ideas. Se produjo un desarrollo en el sector de la industria de la tecnología de la información y son muchos los profesionales de la información involucrados. La recuperación de información pasó a ser utilizada en la Web y es el punto central de las bibliotecas digitales (Saracevic, 1995).

Sin embargo, sigue habiendo algunos problemas:

- El lenguaje utilizado en la indización sigue siendo un punto de estudio. Si bien el lenguaje natural favorece los procesos de indización automática, puede aumentar la llamada y disminuir la precisión de la recuperación de los documentos. Por otro lado, la utilización de un lenguaje de indización aumenta la precisión, pero puede alejarse del lenguaje comprendido y utilizado por los usuarios.
- Casi siempre los buscadores remiten a los títulos de los documentos recuperados, al igual que otros mecanismos se valen de los títulos para la indización y la recuperación de información. Como el título siempre queda bajo la responsabilidad del autor del documento, el control de su coherencia respecto al contenido será siempre posterior a su creación.
- Es imposible pensar en un único modelo para el desarrollo de sistemas de información, y muchas veces el usuario pierde mucho tiempo en la comprensión del funcionamiento de las rutinas que componen el sistema utilizado. Además, otro problema está en el modo en que esos sistemas interactúan con el usuario, a pesar de los avances de las investigaciones sobre la interacción hombre-ordenador.
- Los sistemas automatizados podrían permitir la actualización de los conceptos empleados en lenguajes de indización, permitiendo un registro histórico del empleo de esos conceptos y términos, pero no presentan esa posibilidad, muchas veces a causa del distanciamiento entre el profesional que diseña los sistemas y los profesionales de la información.
- Las revistas científicas poseen normas para orientar y asegurar que los títulos sean coherentes con el contenido. Pero no disponen de herramientas que agilicen ese proceso.

Se han llevado a cabo investigaciones en el área de la ciencia de la computación para la producción de índices y listas que ayuden en el proceso de recuperación de información, principalmente en la Web. En muchos casos se utilizan los títulos como objeto de investigación. En general, la indización de documentos en la Web se da a través del lenguaje natural utilizado en la producción del propio documento, sin que ningún tratamiento previo sea incluido en los documentos involucrados en el proceso. La mayoría de los trabajos en esta área profundizan en cuestiones relativas a la recuperación y al poder de llamada con índices de precisión. Pero estos

trabajos no son interdisciplinarios, no cuentan con la participación de profesionales de la información que conozcan las necesidades de los usuarios, ni siquiera de los profesionales del área abarcada por la investigación. Tampoco existen trabajos que automaticen la coherencia entre los títulos y los contenidos de los textos.

Es necesario desarrollar herramientas de ayuda a la indización que tengan en cuenta la facilidad de indización del lenguaje natural, sin por eso olvidar los beneficios del tratamiento de la información y de los lenguajes de documentación, asegurando que ese proceso produzca resultados que correspondan a la realidad del área.

En este contexto vuelven a destacar los títulos de las publicaciones científicas. Un texto científico posee una estructura compuesta de elementos pretextuales, textuales y postextuales: los pretextuales están compuestos por título, autor del artículo, resumen y palabras clave; los textuales corresponden al texto propiamente dicho, y los postextuales a referencias bibliográficas, resúmenes en otras lenguas, notas al final de texto y anexos (França *et al.*, 2004; Cunha, 2004).

Entre estos elementos, el título siempre ha representado un punto fuerte para la recuperación de información en documentos ya que, por supuesto, es el que representa el contenido que sigue. La recuperación de información en el medio digital se aprovechó también de esa característica, y algunos buscadores privilegian la búsqueda de las palabras requeridas en el título al elaborar la respuesta. Pero, independientemente del medio donde se encuentre el documento, un título no siempre corresponde al contenido. En el caso de textos científicos, las revistas donde generalmente se publican intentan utilizar múltiples normas de orientación a los autores como forma de garantizar la coherencia entre el título y el contenido de los documentos.

En este contexto se suscitó el problema de cómo reconocer, de forma automatizada, si los títulos de artículos científicos son coherentes con su contenido. Una posible solución surgió mediante la comprobación de la aparición de las palabras clave indicadas en los textos por los propios autores en el contenido del título. Esta solución es la que se presenta a modo de experimento en este trabajo.

2.1. Objetivos y metodología

El objetivo del trabajo fue comprobar si era posible verificar la coherencia entre los títulos de los artículos y su contenido de forma automatizada; en concreto, saber si la aparición de palabras clave en los títulos de los artículos constituye un mecanismo aceptable para validar la correspondencia del título de un trabajo científico con su contenido.

La metodología escogida fue experimental, utilizando tecnología de computación para la construcción de un instrumento capaz de verificar esa aparición. Como caso de estudio fueron seleccionados textos científicos en formato electró-

nico, puestos a disposición por las revistas *DataGramaZero* y *Ciência da Informação*, en sus respectivas direcciones electrónicas, en enero de 2005. La elección se hizo: I) por haber sido posible reunir textos de dos colecciones de un mismo período (1999-2004); II) por la uniformidad de los temas tratados en los dos periódicos, incluyendo textos relacionados con las ciencias de la información; III) por la disponibilidad de los textos en formato digital, en su forma integral en la Web, y IV) por el hecho de que ambos periódicos poseen normas de publicación que garantizan la existencia de los títulos y de las palabras clave.

La elección del período al que pertenecían los artículos se debió en primer lugar a la disponibilidad de las revistas, y también al hecho de que se trata de un período privilegiado en cuanto a la producción en ciencias de la información en Brasil, con temas relacionados con la tecnología de la información, las ciencias de gestión y el área del conocimiento, entre otras.

DataGramaZero es un periódico electrónico, bimestral, producido por el Instituto de Adapatação e Inserção na Sociedade da Informação (IASI), una organización no gubernamental, sin vinculaciones políticas, partidistas o religiosas, fundado en noviembre de 1998 y dedicado a los estudios e investigaciones sobre la sociedad de la información. En él, los textos están agrupados por afinidad temática, englobando temas relacionados con las áreas interdisciplinarias de las ciencias de la información, como información y sociedad, información y política pública, información y filosofía o información y comunicación, estando disponibles en versión íntegra en formato HTML.

El periódico tiene unas normas para la publicación de artículos, facilitadas por el propio sitio web. Entre ellas destacan la regla referente a la lengua en que deberán estar los artículos: portugués o español; la exigencia de un resumen temático de unas cien palabras, en portugués (o español) e inglés, título en portugués (o español) e inglés y la exigencia de un conjunto mínimo de cinco palabra clave (*DataGramaZero*, 2004). La primera norma justifica la elección del grupo experimental, formado por todos los textos en lengua portuguesa, lo que resultó en un total de 119 textos, en el período de diciembre de 1999 a octubre de 2004; la segunda exige el resumen en todos los artículos y la tercera es válida para el objetivo del experimento, por la presencia de las palabras clave.

La revista *Ciência da Informação* (2006) es responsabilidad del Instituto Brasileiro de Informação em Ciencia e Tecnología (IBICT). Su primer número data de 1972. Tuvo una periodicidad semestral hasta 1991, cuando pasó a ser cuatrimestral. Es un periódico de gran repercusión en el campo de las ciencias de la información y del sector de información en ciencia y tecnología (C&T), y publica textos de especialistas nacionales y extranjeros. La revista tiene sus artículos indexados o resumidos en *Paschal Thema: Science de L'Information, Documentation*; *Library and Information Science Abstracts*; *PAIS Foreign Language Index*; *Information*

Science Abstracts; *Library Literature*; *Páginas de Contenido: Ciencias de la Información*; *Educación: Noticias de Educación, Ciencia y Cultura Iberoamericanas*; *Referativnyi Zhurnal: Informatika.*

Su contenido incluye artículos, entrevistas, comunicaciones, relatos de experiencias, artículos de sobre literatura y las secciones "in memoriam", "punto de vista" y "editorial". La revista publica trabajos e portugués, español, inglés y francés. Están disponibles en la Web los artículos desde el primer número de 1995. La revista posee normas editoriales, entre las que destacan que el título del trabajo debe "ser breve y suficientemente específico y descriptivo", el resumen debe contener unas doscientas palabras y estar de acuerdo con las normas de la ABNT, así como también deben seguir estas normas la numeración de los títulos, la bibliografía y las citas.

Su elección se justifica por su importancia para el área de las ciencias de la Información en Brasil y por su rigor editorial. De ella fueron escogidos 181 textos en portugués, de diversos tipos, editados entre enero de 1999 y septiembre de 2004.

3. Algo sobre los periódicos científicos

Los periódicos científicos surgieron en el siglo XVII, en Inglaterra, tras la restauración de la monarquía en 1660, como consecuencia de las reuniones que se celebraban durante los años de la guerra civil en distintos locales y ciudades para debatir cuestiones filosóficas.

Nacieron también con el objetivo de facilitar la comunicación entre la comunidad científica. Después de este período de guerra, Londres fue escogida como sede oficial para estas reuniones, que finalmente llevaron a la formación de la Royal Society en 1962. Desde su fundación, esta institución se preocupó por la cuestión de la comunidad científica, influenciada por las ideas de Bacon sobre la posibilidad de una institución científica. Se daba prioridad a la recolección de análisis de las informaciones recibidas, y hay constancia de que sus miembros recorrían países extranjeros buscando datos sobre trabajos llevados a cabo en ellos. Como la tarea era costosa y llevaba tiempo, escogieron como miembros de la Royal Society a personas del extranjero que cumplirían la tarea de comunicar a la entidad, mediante cartas, los trabajos desarrollados en sus países (Meadows, 1999). Solo que la solución fue temporal y el volumen de cartas era tan grande que fue preciso encontrar otra forma de divulgación de los trabajos.

En 1964, Denis de Sallo, un parisino involucrado en esa forma de recogida y difusión de información, comenzó a elaborar un periódico destinado a la publicación de lo que acontecía en Europa. El 5 de enero de 1665 apareció el primer número de *Le Journal des Sçavan*, y con él nació el periodismo científico. En aquella época, según Lemos (1968) y Meadows (1999), el periódico tenía entre sus

objetivos presentar un catálogo de los principales libros aún no publicados en Europa, con informaciones sobre su contenido y su utilidad; incluir necrológicas de celebridades de la época, con bibliografía de sus obras; divulgar experimentos de física, química y anatomía para la explicación de fenómenos naturales, así como describir inventos de máquinas útiles o curiosas; difundir decisiones de los tribunales civiles y eclesiásticos y censuras de universidades; y llevar a los lectores información diversa para alentar la curiosidad humana.

En marzo de 1665 la Royal Society publicó su revista *Philosophical Transactions*, que, al igual que el *Journal des Sçavans*, poseía una amplia cobertura, aunque ambos tomaran caminos más específicos con el paso del tiempo. A partir de allí, el crecimiento de la literatura científica fue exponencial, y la motivación se centra en la necesidad de una difusión y una comunicación eficaces para la comunidad científica, además de ayudar a la integración y la cooperación de los investigadores, contribuyendo también a la legitimación, el reconocimiento del trabajo y la aceptación del investigador en la propia comunidad (Oliveira y Noronha, 2005).

Desde su origen, el periódico científico desempeña la función de registro, de difusión y de institución social. Como registro es un medio formal de control de calidad de la propia revista, además de una fuente de saber científico y de conocimiento público. Como agente de difusión de la información brinda informaciones de interés a la comunidad científica y fomenta la discusión sobre los puntos que en él se exponen. Como institución social otorga prestigio y reconocimiento a los autores, a las instituciones, a los editores y a los evaluadores (Valerio, 1994).

Pero si de un lado están los productores de los documentos, deseosos de que su contenido sea conocido y difundido, del otro lado están los usuarios, con una gran variedad de necesidades. Las unidades de información se sitúan entonces entre ellos, en el vehículo apropiado para la difusión de la información. Y para que este encuentro se lleve a cabo con eficacia, se elaboraron a lo largo de los años procesos de indización y recuperación de información, poniendo gran esfuerzo para perfeccionar la labor de los sistemas de recuperación de información. En líneas generales, los sistemas de recuperación de información se forman a partir del esfuerzo humano y de los procedimientos implantados que buscan facilitar a los usuarios la localización de la información disponible, partiendo de las peticiones realizadas por ellos (Araújo, 1994)

Esos procedimientos hablan de las actividades de representación, almacenamiento, organización y acceso a los documentos (Salton y McGill, 1983). La representación se lleva a cabo a través de procesos de indización, actividad intelectual realizada por profesionales especializados en documentación. El almacenamiento incluye los procesos de gestión de los documentos, independientemente del medio en que se encuentren. La organización, así como la representación, intenta facilitar al usuario el acceso a la información deseada.

3.1. La garantía que viene de las unidades de texto

Hubo algo que llamó la atención durante la elaboración de este trabajo: los textos científicos obedecen a un determinado formato, con normas para su presentación. Esto facilita el tratamiento automatizado mediante programación, favoreciendo el reconocimiento de las partes del texto en los archivos digitales. Puede observarse que algunas partes específicas son prioritarias tanto para quien las escribe como para quien las lee. El título, sin duda alguna, es una parte privilegiada. Como norma, los títulos científicos deben ser explicativos, y dentro de lo posible han de contener datos relevantes sobre el contenido del trabajo; breves, evitando redundancias lingüísticas sin perder la propiedad de presentar el contenido; claros, sin conceptos ambiguos u oscuros; atractivos, que despierten el interés del usuario por su contenido (Ortega, 2003).

Existen normas editoriales regidas por organismos nacionales en todos los países. En el caso de Brasil, hay normas técnicas de la Associação Brasileira de Normas Técnicas (ABNT) que reglamentan la presentación de esta estructura y de las particularidades de cada uno de los elementos. El título debe ser claro y objetivo, pudiendo estar formado por título y subtítulo, y ha de estar siempre en la misma lengua del texto, aunque, si es necesario presentarlo en otras lenguas, puede hacerse a continuación. En la composición de un título de una publicación científica se debe evitar la utilización de paréntesis y fórmulas que dificulten la comprensión del contenido.

Según Krzyzanowski y Ferreira (1998), fue a partir de la década de los sesenta cuando comenzaron a surgir en la literatura estudios sobre evaluación de revistas científicas y técnicas, demostrando la necesidad de que se definieran parámetros de medición que realmente reflejasen la calidad de la información registrada. Los autores hacen referencia a un artículo publicado por Arends (1), donde se relata una evaluación de dos periódicos médicos venezolanos basada en un modelo creado en 1964 por un grupo de trabajo de la UNESCO para la selección de revistas técnicas latinoamericanas. En 1982, Braga y Oberhofer (2) (*apud* Krzyzanowski y Ferreira, 1998) presentaron una propuesta para la evaluación de periódicos brasileros científicos y técnicos, alterando el modelo de la UNESCO. Esos autores proponen un modelo que busca reflejar aspectos de forma de los periódicos, dentro de parámetros medibles.

Lopes Neto *et al.* (2002) presentan una investigación realizada sobre títulos de artículos de un periódico de enfermería. Este trabajo merece atención, ya que el problema que exponían se enfocaba a la adecuación de los títulos de las investigaciones de enfermería para traducir lo que realmente había sido estudiado. En general, los autores llegaron a las siguientes conclusiones:

- De manera general, la economía de palabras era una condición para la objetividad y la veracidad de los títulos.
- Los títulos debían presentar las contribuciones a los lectores de investigaciones científicas de manera directa, facilitando su búsqueda sin generar ansiedad o frustración.
- Los títulos exuberantes, elaborados intentando mostrar un lenguaje filosófico y científico, comúnmente dejan de expresar el contenido del trabajo. El uso de un lenguaje rebuscado puede decepcionar a los lectores, cuando lo ideal sería motivarlos.
- La mayoría de los artículos analizados presentaban títulos considerados como adecuados de acuerdo con los criterios de análisis utilizados, aunque en una cantidad significativa de artículos los títulos con una captación solo adecuada en parte, ya que, según la percepción de los autores, no coincidían con el contenido expresado en el cuerpo del trabajo.
- El empleo de metáforas en los títulos fue verificado en solo el 7,35% de los casos y el pleonasmo en el 5,38%.
- Se encontraron títulos largos en el 16,37% de los artículos, y títulos con problemas de puntuación (empleo inadecuado de signos de puntuación, como interrogación, dos puntos, etcétera), en el 13,24%.
- En relación con la claridad del título, se observó en el 66,17% de los casos.
- Las palabras clave utilizadas en los artículos fueron captadas de forma parcial en los títulos y, según estos autores, se debe prestar una mayor atención a esta cuestión.

Ortega (2003) recuerda que un título debe ser atractivo, pues es el primer elemento que el usuario busca en una pesquisa bibliográfica. Y, algunas veces, se dejan de lado trabajos científicos debido a la poca atracción que ejercen sus títulos, o a que tienen títulos muy cortos o muy largos, que hacen dudar en el momento de escoger el documento. Para este autor, el título deberá presentar siempre variables en consonancia con el estudio realizado y, si es posible, incorporar datos del momento y el lugar donde se realizó la investigación. Se deben evitar las abreviaturas, las fórmulas y los signos de interrogación. En este último caso, hace hincapié en que en el título no se presenta el problema o la cuestión suscitada en el documento científico.

La revista *DataGramaZero* tiene normas para la presentación de artículos, pero entre ellas no existe ninguna específica para la reglamentación de los títulos. La revista *Ciência da Informação* tiene entre sus normas un párrafo destinado a los títulos de publicaciones. De acuerdo con este, el título debe ser breve, específico y descriptivo, e incluir las palabras clave que representen el contenido del texto. Esta regla será observada en la herramienta elaborada en este trabajo.

4. El experimento

Para este trabajo se desarrollaron programas que utilizan un lenguaje PHP (*hypertext preprocessor*), en su versión 4. Se trata de un lenguaje orientado al desarrollo de aplicaciones web. La elección se debió principalmente a la posibilidad que ofrece para la administración de *string* ('serie de letras'), lo que facilita la realización del proceso. Para el almacenamiento de los datos se optó por el administrador de base de datos Mysql, software libre que utiliza un lenguaje SQL (*structured query language*) como lenguaje de manipulación de datos (DML: *data manipulation language*). Una de las principales características del Mysql es su total integración con el PHP. La interfaz utilizada para consultas en el proyecto fue el EasyPHP (EasyPHP, 2005).

Los tipos de documentos escogidos para el experimento fueron textos científicos: artículos, recensiones, comunicaciones, relatos de experiencias y artículos sobre literatura de dos periódicos brasileros disponibles en la Web. Esta opción se basó en el la posibilidad de identificar en ese tipo de textos la unidad de texto elegida, el título, apoyándonos en los trabajos de França *et al.* (2004) y Cunha (2004) y en las normas técnicas de la ABNT.

Para cada compilación se creó un directorio y los artículos se guardaron en su compilación específica. Todos los textos, originalmente en formato HTML, se convirtieron a formato de texto (TXT), y los nombres externos de los archivos se uniformizaron: artículo + año + mes + número del artículo. txt. Así, el archivo *artículo1999ene1.txt* se refiere al texto número 1 de la edición de enero de 1999. Se optó por no colocar el nombre de la compilación en el nombre externo de los archivos, ya que los textos se mantendrían en directorios diferentes para cada compilación.

Se consideraron atributos calificadores de cada compilación la fuente de donde fueron extraídos los documentos y el período de producción de los mismos, para facilitar posteriormente la recuperación de datos estadísticos por esos cortes.

En el experimento realizado se siguieron los siguientes pasos: agrupación de los artículos de cada compilación en un directorio, ya convertidos a formato TXT; elaboración de programas para la extracción de los títulos y las palabras clave de cada artículo, dando lugar a la tabla "artículo" y la tabla "palabra clave"; generación de datos estadísticos de la aparición de palabras clave en los títulos; clasificación de los artículos cuyas palabras clave no existían en el título; análisis cuantitativo y cualitativo de los datos. La figura 1 ilustra estos procedimientos.

La creación de las tablas en el administrador de base de datos se produjo a partir del modelo de entidades y relaciones (Elmasri y Navathe, 2000) representado en la figura 2.

La tabla "artículo" contiene las columnas "número de artículo", "título del artículo", "nombre externo del archivo", "año de publicación". La tabla "palabra

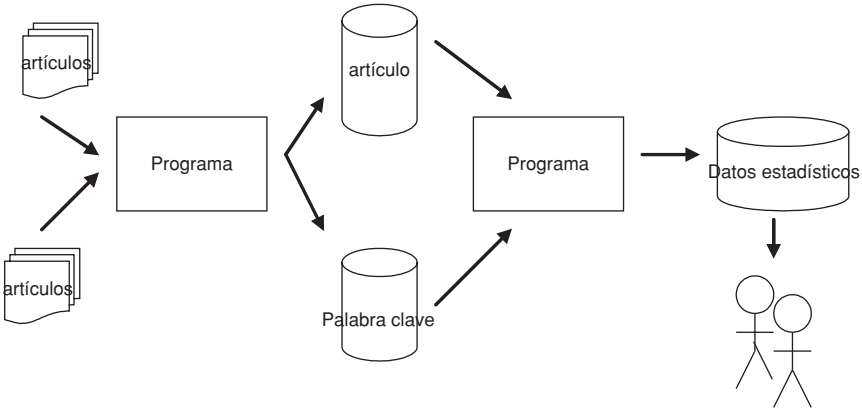

Figura 1. Experimento realizado

Figura 2. Modelo de entidades y relacionamientos para las bases artículo *y* palabra clave

clave" se creó con las columnas "código de la palabra clave" y "palabra clave". El código de la palabra clave fue una secuencia numérica. Para facilitar futuras verificaciones, la palabra clave se consideró un atributo único. La relación "artículo" x "palabra clave" informa el número de apariciones de la palabra clave en el título del artículo.

La muestra se componía de 300 textos científicos, todos en lengua portuguesa, publicados entre 1999 y 2004, en dos grupos. La tabla 1 presenta el total de textos clasificados por grupo y año de publicación, datos ilustrados por la figura 3.

En porcentajes, el 60% de los textos pertenecen a la revista *Ciência da Informação* y el 40% a la revista *DataGramaZero.* Si se clasifican por año, la frecuencia de artículos tiene un crecimiento lineal, siendo los años 2002, 2003 y 2004 los que cuentan con mayor número de textos aparecidos.

La base de palabras clave se componía de 957 palabras clave. De ellas, 475 se originaron a partir de textos de la revista *Ciência da Informação*, 306 de la revista

Año	Ciência da Informação	DataGramaZero
1999	32	5
2000	26	16
2001	24	23
2002	31	24
2003	36	23
2004	32	28
Total	181	119

Tabla 1. Número de artículos por año en cada grupo

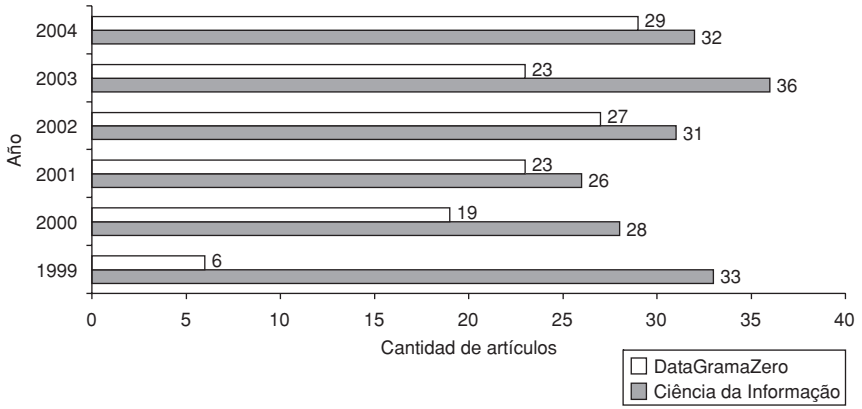

Figura 3. Número de artículos por año en cada grupo

DataGramaZero y 176 de las dos revistas simultáneamente. La figura 4 representa esas cifras en porcentajes.

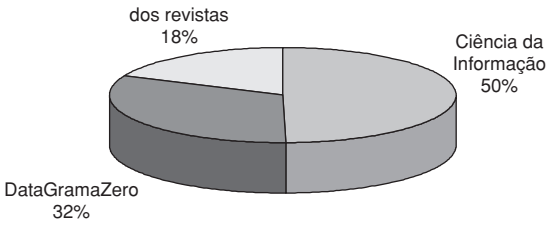

Figura 4. Distribución porcentual de la aparición de palabras clave en los títulos de los grupos

Entonces se generaron y almacenaron en bases de datos las estadísticas de aparición de las palabras clave en los títulos de los artículos. En la revista *Ciência da Informação* se encontraron 137 artículos, con alguna palabra clave en su título, lo que equivale al 76% de los artículos de este grupo. En la revista *DataGramaZero* se hallaron 94 artículos con alguna palabra clave en su título, es decir, el 79% de este grupo.

No fue posible saber cuándo la revista *Ciencia da Informação* incorporó es sus normas de publicación la recomendación de que existieran palabras clave en los títulos. En esta revista, 44 artículos no tenían ninguna de sus palabras clave en el título, lo que corresponde al 24% de los artículos de este grupo. La figura 5 presenta una distribución de los artículos donde no fueron encontradas las palabras clave por año de publicación.

Como continuación del experimento, se separaron los 69 artículos cuyas palabras clave no aparecían en el título y se comprobó si sus palabras clave aparecían en otras unidades de texto elegidas: resumen, cuerpo del texto, referencias bibliográficas. También se verificó si las palabras del título existían en las unidades de texto de estos artículos.

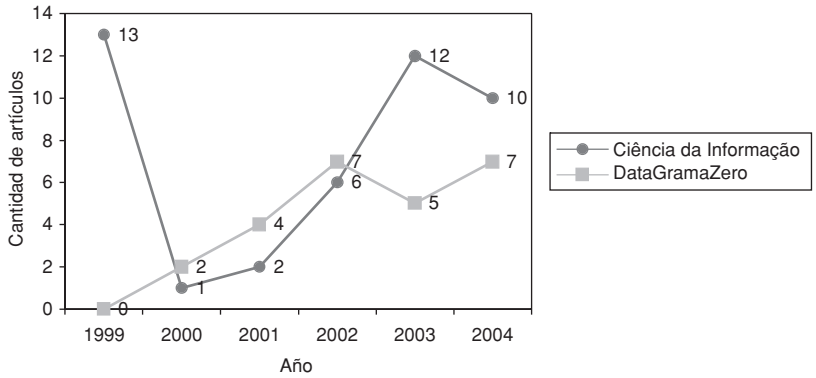

Figura 5. Número de artículos sin aparición de palabra clave por año/grupo

4.1. Análisis de los datos

Los artículos que tenían sus palabras clave en el título fueron considerados como artículos de títulos coherentes con su contenido. Se analizaron estos artículos y esta hipótesis se confirmó: no se encontraron en este grupo artículos cuyo título fuese incoherente con el tema. Pero, curiosamente, se halló un artículo con una mayor creatividad en la elaboración del título: "O unicórnio (o rinoceronte, o ornitorrinco...), a análise documentária e a linguagem documentária". Las palabras

clave eran *análise documentária*, *linguagem documentária*, *informação documentária*, *terminología*, *lingüística*, *semiótica*.

El análisis del texto confirma que la autora utilizó la creatividad como una forma de llamar la atención sobre el artículo. El texto tiene como punto de partida el capítulo "Marco Polo y el unicornio" del libro *Kant y el ornitorrinco*, de Umberto Eco, y la autora realiza una lectura exploratoria creando un paralelismo entre el proceso del conocimiento y el de la representación de la información documentaria.

Se analizaron los artículos cuyas palabras clave no existían en el título y se constataron los siguientes puntos:

a. La no aparición de las palabras clave, en algunos casos, está motivada por diferencias de grafía, ya sea entre singular y plural o entre construcciones lingüísticas distintas. Las diferencias de presentación singular/plural entre palabras del título y palabras clave se dieron en 14 artículos, el 20,29% de los casos en los que no aparecían las palabras clave en el título. Son ejemplos de esto:

   Título: *Webmuseus: aparatos informacionais...*
   Palabras clave: *webmuseu*, *aparato informacional...*
   Título: *Biblioteca digital de...*
   Palabras clave: *bibliotecas digitais...*

   Podría considerarse que estos artículos tienen las palabras clave en el título, pero se escogió presentarlos aquí para resaltar que en experimentos futuros relativos a la coincidencia de palabras debe tenerse en cuenta la cuestión de la diferencia entre plural y singular.

   En cuanto a las distintas construcciones lingüísticas y el empleo de sinónimos, a veces los autores escogen grafías diferentes para las palabras del título y las palabras clave. Esto no representa ausencia de representación de las palabras clave en el título. Sin embargo, la utilización de filtros o inteligencia en la programación durante la comparación de palabras aumentaría la complejidad del experimento, exigiendo un diccionario muy completo para seleccionar la posibilidad de sinónimos para las palabras. Lo ideal es que los autores actúen siempre del mismo modo en la construcción lingüística entre títulos y palabras clave. Este hecho fue observado en 7 artículos, el 10,14% de los casos, en cuyos títulos no aparecían palabras clave. Ejemplifican el hecho los casos siguientes:

   Título: ... *uma abordagem bibliométrica*
   Palabras clave: *bibliometria...*
   Título: ... educação e comunicação...
   Palabras clave: *comunicação e educação*

Se analizaron los artículos contenidos en este grupo y no se encontró ninguna inconsistencia entre título y contenido.

b. Hay casos de títulos específicos con indicación de palabras clave genéricas o viceversa. Se trata de una elección de los autores y ha de respetarse. Para los procesos de recuperación de información es importante que palabras clave y títulos coincidan de alguna manera, y la elección de este tipo de construcciones debe implicar que haya sintonía entre estas palabras. Los términos utilizados en las palabras clave deben al mismo tiempo seguir normas para localizar el tema del texto de una manera más abarcadora, pero también deben dar facilidades a los lectores en sus *especificidades*. Generalmente, el número de palabras clave recomendado por los periódicos gira en torno a cinco, suficiente para dar noción del conjunto y del objeto específico. Ese punto fue observado en 12 artículos, el 17,39% de los casos. Son ejemplo de estas apariciones:

> Título: *Áreas do conhecimento*
> Palabras clave: *organização do conhecimento*, *representação do conhecimento*, *classificação do conhecimento*, *classificação em ciência e tecnología*, *classificação de projetos de pesquisa e desenvolvimento*
> Título: *A formação profissional no século XXI: desafios e dilemas*
> Palabras clave: *educação dos bibliotecários*, *profissional da informação*

En estos artículos no se halló ninguna inconsistencia entre títulos y contenido.

c. El empleo de precoordinación en las palabras clave es un hecho común y fácil de observar en textos científicos. Aunque los autores en la mayoría de las veces dominen el universo del discurso de su área de actuación, existe una tendencia a precoordinar las palabras clave en un intento de hacerlas coincidir con el contenido del artículo. El orden en esto que se hace muchas veces es diferente del orden en que se colocan las palabras en los títulos, lo que impide la verificación de la existencia de palabras clave en el título. Además, hay palabras clave que en la recuperación se utilizan con medios de poscoordinación. Sería ideal que las revistas orientaran a los autores en este sentido. El hecho fue observado en 32 artículos, el 46,38% de los casos, y ninguno de ellos presentó inconsistencia entre título y contenido.

d. Títulos más creativos, con empleo de metáforas o de palabras con doble sentido. Ese tipo de construcción está contraindicado por Ortega (2003), França *et al*. (2004) y Cunha (2004), y es desaconsejado por las normas de algunas revistas científicas. Pero es elección de los autores. En este caso se encontraron 4 artículos, el 5,80% de los casos. Aunque las palabras clave de estos artículos sean coherentes con el contenido, la ausencia en el título se debe a la construcción escogida:

Título: *O olhar da consciência possível sobre o campo científico*
Palabras clave: *teoría da ciencia da informação*, *sociología da informação*, *historia da ciencia da informação*, *comunicação científica*, *responsabilidade social*
Título: *A explosão do filósofo e a obsessão de informação*
Palabras clave: *explosão da informação*, *crecimiento do conhecimento*, *sociedade da informação*, *obsessão social*
Título: *A libertade das vozes*
Palabras clave: *libertade da informação*, *tecnología e inovação*, *historia da informação*, *geração de Conhecimento*
Título: *A informação em seus momentos de passagem*
Palabras clave: *criação da informação*, *contexto da informação*, *gerenciamento da informação*, *agentes inteligentes*, *analise textual*, *ciencia da informação*

Como alternativa al análisis de los datos, se buscó el resultado del procesamiento de las palabras clave en las otras unidades de texto (resumen, cuerpo del texto y referencias bibliográficas) para los artículos en cuyos títulos no había ninguna de sus palabras clave. Se verificó que en el 13% de los casos (9 textos) no se halló ninguna de las palabras clave en el cuerpo del texto, en el 36% (25 textos) no se encontraron en los resúmenes y, en el 58% (40 textos), en las referencias bibliográficas.

Entonces se asignó una categoría para el análisis de estos artículos obedeciendo a una jerarquía de prioridades entre resumen, cuerpo del texto y referencias bibliográficas, a través de los siguientes criterios: a) artículos donde las palabras clave aparecían en el resumen; b) artículos donde las palabras estaban en el cuerpo del texto y no en el resumen; c) artículos donde las palabras clave se encontraban en las referencias bibliográficas y no en el resumen ni en el cuerpo del texto; d) artículos donde las palabras no aparecían ni en el resumen ni en el cuerpo del texto ni en las referencias bibliográficas. Estos datos se muestran en la tabla 2. En ella, si el artículo posee una de sus palabras clave en el resumen, independientemente de que la tenga también en el cuerpo del texto o en las referencias bibliográficas, se incluyó solo en el apartado del resumen.

El análisis de los artículos mediante este agrupamiento no sumó nuevos puntos de estudio a los que ya se conocían anteriormente: grafías diferentes, títulos es-

Categoría de aparición de las palabra clave	Total de artículos
Al menos en el resumen	44
Al menos en el cuerpo del texto	16
Al menos en las referencias bibliográficas	1
Ni en el resumen ni en el texto ni en las referencias bibliográficas	8

Tabla 2. Aparición de las palabras clave en los artículos que no las tienen en los títulos

pecíficos con palabras clave genéricas o viceversa, precoordinación de palabras clave y exceso de creatividad en los títulos. El mayor problema se encuentra en la elección de las palabras clave y no en los títulos de los artículos. Las revistas científicas han de prestar atención a la orientación de los autores en este sentido. Los trabajos futuros deben tener en cuenta de algún modo los problemas que presenta el uso de grafías diferentes: plural y singular, masculino y femenino, etcétera.

La verificación de las palabras del título en las unidades de texto dio como resultado que 22 artículos incluían palabras que solamente aparecían en el título. El análisis de estos artículos mostró que este puede ser otro camino, pero en el caso de las compilaciones utilizadas aquí no había títulos incoherentes con el contenido.

5. Conclusión

Se escogieron artículos de dos revistas científicas disponibles en la Web. Se verificó la aparición de palabras clave en los títulos de estos artículos. Algunos de ellos no tenían las palabras clave en el título. El problema reside más en la elección de las palabras clave que en la incoherencia entre títulos y contenidos. Solo 4 artículos, en cuyos títulos no aparecían las palabras clave designadas, poco más del 1% de la muestra, presentaban algún conflicto entre los títulos y los contenidos, debido a la creatividad de los autores.

Después del experimento se realizó un análisis cualitativo de la muestra y se comprobó que no existían títulos incoherentes con el contenido. Lo más destacado de este análisis fueron los títulos ya mencionados como más creativos. Se confirma que, de los artículos con problemas de inconsistencia entre texto y título, los datos cuantitativos aquí presentados confirman la existencia de este problema.

Reconocer títulos no coherentes con el contenido de forma automática es un problema que se ha de resolver. Las palabras clave pueden ser un camino. Se debe observar la presencia de palabras clave en los títulos de los artículos durante el proceso de selección para la verificación del título junto al contenido del artículo. Se cree también que las revistas digitales podrían tener vocabularios controlados elaborados a partir de estas palabras. Esto también facilitaría a los autores la elección de títulos y palabras clave. Las actuaciones en este sentido facilitarían la búsqueda y el encuentro de documentos ante las necesidades de los usuarios.

Notas

(1) Arends, L. (1968). Las revistas médicas venezolanas: evaluación de su calidad. // Acta Científica Venezolana. 19 (1968) 145-151.

(2) Braga, G. M.; Oberhofer, A. (1982). Diretrizes para avaliação de periódicos científicos e técnicos brasileiros. // Revista Latina de Documentación. 2:1 (ene.-jun. 1982) 27-31.

Referencias

Araújo, Vânia M. R. H. (1994). Sistemas de recuperação da informação – SRIs. // Sistemas de recuperação da informação: nova abordagem teórico-conceitual. Tesis doctoral. Río de Janeiro: Escola de Comunicação, Universidade Federal do Rio de Janeiro, 1994. 84-122.

Bush, V. (1945). As we may think. // The Atlantic Monthly. 176:1 (jun. 1945) 101-108.

Castells, Manuel (1999). A revolução da tecnologia da informação. // A sociedade em rede, vol. 1. São Paulo: Paz e Terra, 1999. 49-86.

Ciência da Informação. X (ago. 2004). URL: <http://www.ibict.br/cienciadainformacao>. Consultado: 2004-12-09.

Cunha, H. R. S. (2004). Padrão PUC Minas de normalização: normas da ABNT para apresentação de artigos em periódicos científicos. URL: <www3.pucminas.br/documentos/normalizacao_artigos.pdf>. Consultado: 2005-02-27.

DataGramaZero. Revista de Ciência da Informação. URL: <http://www.dgz.org.br/>. Consultado: 2004-12-09.

Dodebei, Vera Lúcia Doyle (2002). Tesauro: linguagem de representação da memória documentária. Niterói: Intertexto, 2002.

EasyPHP. URL: < http://www.easyphp.org/>. Consultado: 2005/06/01.

Elmasri, R.; Navathe, S. B. (2000). Fundamentals of database systems. Reading, Mass.: Addison-Wesley, 2000.

Foskett, Anthony Charles (1973). A abordagem temática da informação. São Paulo: Polígono, 1973.

França, J. L., *et al.* (2004). Manual para normalização de publicações técnico-científicas. Belo Horizonte: UFMG, 2004.

Krzyzanowski, R. F.; Ferreira, M. C. G. (1998). Avaliação de periódicos científicos brasileiros. // Ciência da Informação. 27:2 (may.-ago. 1998) 165-169.

Lemos, Antônio Agenor Briquet de (1968). Presente e futuro do periódico científico. // Correio Braziliense. Caderno Cultural. 13 jul. (1968) 3. URL: <http://www.briquetdelemos.com.br/editor1.htm>. Consultado: 2005/10/20.

Lopes Neto, D. *et al.* (2002). Análise de títulos de artigos de pesquisas publicadas em um periódico brasileiro de enfermagem. // Revista Latino-Americana de Enfermagem. 10:1 (jan.-fev. 2002) 77-84.

Meadows, A. J. (1999). A comunicação científica. Brasília: Briquet de Lemos, 1999.

Murriel, Gatti (1998). ¿Por qué prestar atención al lenguaje? // Boletín Informativo de Temas Lingüísticos del Departamento Académico de Humanidades de la Universidad del Pacífico. 1:1 (jul. 1998). URL: <http://www.up.edu.pe/coine/Boletin1/TRASFOND.HTM>. Consultado: 2005/06/10.

Oliveira, Érica B. P. M. de; Noronha, Daisy P. (2005). A comunicação científica e o meio digital. // Informação & Sociedade: Estudos. 15:1 (2005). URL: <http://www.informacaoesociedade.ufpb.br/>. Consultado: 2005-10-25.

Ortega González, Javier (2003). El título en las publicaciones científicas: algunos consejos para su estructuración. // Revista Médica IMSS. 4:41 (jul.-ago. 2003) 355-358.

Salton, Gerard; McGill, Michael J. (1983). Introduction to modern information retrieval. New York: McGraw Hill, 1983.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. // Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle: ACM Press, 1995. 138-146.

Valério, Palmira Moriconi (1994). Espelho da Ciência: avaliação do Programa Setorial de Publicações em Ciência e Tecnologia da FINEP. Brasília: FINEP/IBICT, 1994.