UNIVERSIDAD NACIONAL DE SAN LUIS
FACULTAD DE CIENCIAS FÍSICO
MATEMÁTICAS Y NATURALES

Tesis de Maestria en Ciencias de la Computacion

Título del Trabajo

Carlos Guillermo Velázquez

Director: Dr. Marcelo Luis Errecalde
Co-Director: Dra. Leticia Cagnina

San Luis Argentina

2015

Prefacio

Sección Vacía

Carlos Guillermo Velázquez
carvear20@yahoo.com.ar
Universidad Nacional de San Luis
San Luis, [fecha presentación].

Resumen

Sección Vacía

Índice general

Capítulo 1

Introducción

Las bases de datos tradicionales son construidas basándose en el concepto de búsqueda exacta: la base de datos es dividida en registros y cada registro contiene campos completamente comparables. Las consultas a la base de datos retornan todos aquellos registros cuyos campos coinciden exactamente con los aportados en tiempo de búsqueda (búsqueda exacta).

A través del tiempo las bases de datos han evolucionado para incluir la capacidad de almacenar nuevos tipos de datos tales como imágenes, sonido, video, etc. Estructurar este tipo de datos en registros para adecuarlos al concepto tradicional de búsqueda exacta es difícil en muchos casos y hasta imposible si la base de datos cambia más rápido de lo que se puede estructurar (como por ejemplo la web). Aún cuando pudiera hacerse, las consultas que se pueden satisfacer con la tecnología tradicional están limitadas en variaciones de la búsqueda exacta.

Nos interesan las búsquedas en donde se puedan recuperar objetos similares a uno dado. Este tipo de búsqueda se conoce con el nombre de búsqueda por similitud, y surge en diversas áreas; tales como reconocimiento de voz, reconocimiento de imágenes, compresión de texto, biología computacional, entre otras.

Todas estas aplicaciones tienen algunas características comunes. Existe un universo X de objetos y una función de distancia $d : X \times X$ que modela la similitud entre los objetos. El par (X, d) es llamado espacio métrico REF. La base de datos es un conjunto UX, el cual se preprocesa a fin de resolver búsquedas por similitud eficientemente.

Una de las consultas típicas en este nuevo modelo que implica recuperar objetos similares de una base de datos es la búsqueda por rango, que denotaremos con $(q, r)d$. Dado un elemento de consulta q, al que llamaremos query y un radio de tolerancia r, una búsqueda por rango consiste en recuperar aquellos objetos de la base de datos cuya distancia a q no sea mayor que r.

Otra consulta que se utiliza para recuperar objetos similares es la búsqueda de los k-vecinos más cercanos, donde tenemos el elemento de consulta q y la cantidad de objetos de base de datos que se quieren recuperar, que llamaremos

k. La cercanía de estos k objetos está dada por el valor de la función de distancia d hasta la query q.

Los cómputos realizados durante una búsqueda implican cálculos de la función de distancia d, operaciones adicionales (sumas, restas, comparaciones, etc) y tiempo de I/O si el índice y/o los datos se encuentran en memoria secundaria. Las búsquedas por similitud pueden ser resueltas trivialmente por medio de una búsqueda exhaustiva, con una complejidad $O(n)$. Para evitar esta situación, se preprocesa la base de datos por medio de un algoritmo de indexación con el objetivo de construir una estructura de datos o índice, diseñada para ahorrar cálculos en el momento de resolver una búsqueda.

Básicamente existen dos enfoques para el diseño de algoritmos de indexación en espacios métricos: uno está basado en Diagramas de Voronoi [1, 6, 2, 7] y el otro está basado en pivotes [6, 3, 4, 5]. En este trabajo abordamos el estudio de algoritmos de indexación basados en pivotes, enfocándonos en una aplicación que utilice este tipo de algoritmos en el ámbito del comercio electrónico.

Los algoritmos basados en pivotes construyen el índice basándose en la distancia de los objetos de la base de datos a un conjunto de elementos preseleccionados llamados pivotes.

1.1. Objetivos

El comercio electrónico, también conocido como e-commerce consiste en la compra y venta de productos o de servicios a través de la web. El desarrollo de nuevas tecnologías ha permitido que la capacidad y volumen de las comunicaciones se expanda de una manera exponencial, esto ha facilitado que el comercio electrónico tenga también un crecimiento exponencial.

Para el desarrollo de un sitio de comercio electrónico hay varios problemas que deben resolverse tales como administración de categorías de productos, búsqueda de productos, encriptación de datos, registración de usuarios, administración de medios de pago, entre otros. En este trabajo nos hemos centrado específicamente en el problema de búsqueda de productos.

Nuestra meta es utilizar búsquedas por similitud sobre las descripciones asociadas a los productos con el fin de estudiar el desempeño de los algoritmos basados en pivotes en un caso real.

1.2. Como se organiza la tesis

Hemos organizado este informe en dos partes: En la primera parte se introducen los conceptos necesarios para comprender el informe y en la segunda parte presentamos el trabajo realizado y los resultados que obtuvimos al aplicar los algoritmos de indexación.

1.2.1. Primera Parte

Capítulo 1: Definimos la motivación y los objetivos del trabajo.

Capítulo 2: Presentamos los conceptos básicos de las búsquedas en espacios métricos con el objetivo de dar el marco teórico necesario para comprender y fundamentar el desarrollo realizado en este trabajo final. También explicamos la situación actual del comercio electrónico, sobre el cual vamos a aplicar los conceptos teóricos.

Capítulo 3: Definimos completamente en qué consiste la aplicación de espacios métricos a esta base de datos real de comercio electrónico.

1.2.2. Segunda Parte

Capítulo 4: Describimos el trabajo realizado. Presentamos la implementación de dos algoritmos de indexación basados en pivotes: Selección de pivotes en forma random y selección de pivotes en forma incremental. También detallamos las diversas variaciones implementadas a fin de encontrar la que mejor se adapte al tipo de búsqueda que queremos realizar.
Capítulo 5: Mostramos la evaluación experimental de las variaciones implementadas.
Capítulo 6: Presentamos las conclusiones obtenidas como así también algunos puntos interesantes para abordar en futuros trabajos.

Capítulo 2

Calidad de la Información

La *calidad* tomada como su forma más general, se define como la propiedad o conjunto de propiedades inherentes a algo, que permiten juzgar su valor. De manera más particular, la información que pretendemos que se encuentre en los documentos, debe cumplir con ciertos atributos que denotan la calidad de los documentos.

La determinación de las características de calidad que ellos deben poseer puede variar dependiendo de los objetivos y el recurso tratado, sin embargo, algunas de ellas son comunes a todos los documentos como lo son: contenido apropiado, actualización, exactitud y accesibilidad entre otros.

La aparición de algunas o todas las características en los documentos trae aparejado un conjunto de ventajas tales como la mejora y precisión del contenido.

Organización del capítulo. La subsección 2.1 presenta una introducción general al concepto de calidad. La subsección 2.2 se enfoca en mostrar aspectos de calidad relacionados a la Web. En la subsección 2.3 se introducen conceptos y características que son considerados de calidad en los artículos de Wikipedia. Al mismo tiempo se presentan los diferentes niveles de calidad existentes en dichos artículos.

2.1. Calidad de la información general

2.2. Calidad de la información en la Web

Que *"la información es poder"*, se sabe desde hace mucho tiempo. Hoy en día, las demandas de información aumentaron pasando a ser un bien necesario en nuestra vida diaria que debe ser actualizado constantemente para no quedar desfazado en el tiempo. La información puede ser definida como el mensaje usado para representar un hecho con el fin de incrementar el conocimiento y el aprendizaje, los cuales son cada vez más necesarios para desarrollar cualquier actividad.

Cuando se hace uso de Internet para recoger información, es necesario definir

exactamente lo que se busca, que tipo de información se necesita, cual es el objetivo de la búsqueda y decidir fundamentalmente que fuentes son las más confiables para poder satisfacerla, sobre todo teniendo en cuenta la gran variedad de material que se puede encontrar en la Web.

Cada día, surge nueva información que es fácil y gratuitamente publicada en la Web, la cual se ha convertido en un potente medio de comunicación y difusión en el que cualquier persona, organismo o empresa que disponga de una computadora y una conexión a Internet, puede generar información en la misma. Sumado a esto, se eliminan todas las barreras tradicionales de los medios de edición impresa, donde para poder publicar se requiere pasar por un proceso de evaluación, de filtrado y de revisión, y cumplir con las normas de publicación propias de cada editorial.

Debido a la propia naturaleza de la red, dinámica, a menudo desorganizada y con información dudosa o desconocida, la calidad de la información ha sido desde sus orígenes cuestionada. En contraposición, nos encontramos con la importancia que en el ámbito de las Ciencias de la Documentación y de la Información ha tenido siempre la evaluación tanto de la calidad de la información como de los sistemas de recuperación de información, con el objetivo de proporcionar al usuario la información que necesita y que además sea útil y de calidad. Como consecuencia directa, detectamos que los usuarios se enfrentan a unos nuevos retos como son localizar información de calidad y útil en la red, y evaluar dicha información para verificar su calidad.

Queda en claro que la calidad de la información actualmente es un concepto multidimensional que combina diversos factores [WS96]. Una interpretación ampliamente aceptada de la calidad de la información es la "aptitud para el uso en una aplicación práctica" [WS96]. De manera similar, Juran y Godfrey [JG99] interpretan la información como "de calidad si son aptas para los usos previstos en la operación", es decir, dependiendo directamente del contexto en el que se aplica.

Otra definición utilizada es la que sostiene que la calidad de la información de un recurso será determinado por la capacidad de satisfacer las necesidades de información de la persona que lo utilice o consulte, sin embargo, esto puede ser muy relativo, ya que la apreciación de la calidad es muy subjetiva, siendo para algunas personas de calidad lo que para otras no lo es. Lo que si podemos afirmar es que la calidad de la información no es una realidad medible de forma exacta y unívoca sino que puede percibirse de diferentes dimensiones, y la cantidad de dimensiones que se tomen en cuenta determinará la exhaustividad de su estudio.

En un principio, la diferencia entre calidad de datos e información de calidad definía a la primera como "la aptitud para el uso", es decir, la capacidad de una colección de datos para satisfacer requisitos del usuario [Str97, Cap02]. Por otro lado, calidad de datos es un concepto multidimensional [Cap02]. En la literatura se han atribuido categorías y dimensiones para encarar problemas relacionados con la calidad de datos e información de calidad. El marco mas utilizado es el

propuesto por [Str97], el cual plantea una clasificación basada en la perspectiva del usuario como se muestra en el cuadro 2.1.

Categorías	**Dimensiones**
Intrínseca	Exactitud, Objetividad, Credibilidad, Reputación
Accesibilidad	Accesibilidad y Seguridad
Contextual	Relevancia, Valor añadido, Oportunidad, Completitud, Información Cantidad
Interpretabilidad representacional	Fácil entendimiento, Representación concisa, Representación coherente

Cuadro 2.1: Categorías y dimensiones para la calidad de la información y datos de calidad.

Actualmente, no existe todavía un consenso acerca de la distinción entre datos de calidad e información de calidad. Madnick [SEMZ09], determina el uso de calidad de datos refiriéndose a detalles técnicos como la integración de registros de diferentes bases de datos, mientras que utiliza calidad de información para referirse a detalles no técnicos como la importancia de la información para cierto usuario, además de ofrecer una visión global de las dimensiones de la calidad, incluyendo precisión, confiabilidad, etc. Por otra parte, los teóricos de la información, utilizan el término *dato* para referirse a textos planos que necesitan ser procesados para convertirse en información útil, mientras que *calidad de la información* es utilizado para referirse al resto de los casos. En un principio, las investigaciones acerca de información de calidad y datos de calidad surgió en los sistemas de información [Str97, Lee02]. Luego se extendió a diferentes contextos, tales como sistemas cooperativos [Fug02, Mar03], almacenes de datos [Bou01, Zhu] y comercio electrónico [Kat99] entre otros.

La calidad de la informacion en Internet particularmente, requiere de un proceso continuo de planificación, análisis, diseño, implementación, promoción e innovación, con el fin de asegurar que se cubran las necesidades de los usuarios en cuanto a contenido, presentación y usabilidad. Sin embargo, debido al esfuerzo que requiere esta tarea sumado a la naturaleza caótica de Internet, que dificulta la búsqueda, identificación y localización de la información deseada, muchas veces se deja de lado y hace que encontremos en Internet recursos de diferentes calidades. Pero el beneficio a largo plazo es mayor en términos de prestigio, marketing, difusión del conocimiento, aunque es bien conocido por las personas o entidades que apuesten por la calidad de la información que ésta tiene un precio, un costo en términos económicos (tiempo que se tarda en publicar es mayor, las revisiones

y mejoras requieren tiempo y personal). Estas razones nos llevan a que si queremos ofrecer información de calidad tengamos en cuenta que es un proceso de mejora constante, que implica llevar a cabo varias acciones, como por ejemplo:

- Uso de checklists o listas para la evaluación de la información.
- Hacer uso de la opinión de los usuarios acerca de la información proporcionada.
- Supervisión y control de la información que se publica.
- Controles de calidad exhaustivos antes de la publicación.

Los documentos electrónicos, constan de dos componentes la forma y la información. A pesar de que ambos son muy importantes, los usuarios, en general, se interesan más en el contenido que en la forma en que se presenta el documento. Sirve de poco que la información este completa y mal organizada y viceverza [Mol].

El boom de la Internet trajo aparejado una creciente diversidad de contenidos dejando en evidencia que la evaluación de los documentos Web generados son un punto más que importante. Esto permitió la introducción de métricas de calidad en los diferentes enfoques de recuperación de la información como ranking sesgado, distancia de la colección de los documentos y radio de ruido de la información, entre otros [MBD11, PT10, ZC05, ZG00]. Cabe destacar que no existe un indicador perfecto que nos mida la calidad o el valor de la información. Algunos de ellos son mas fáciles de medir que otros ya que se encuentran muy relacionados entre si. [R.97] sostiene que determinar la calidad de la información es un arte, ya que hay que inferir a partir de un conjunto de indicadores, basados en la utilidad o propósito con el que se quiera utilizar la fuente de información. De manera general, actualmente, pocas fuentes de información van a satisfacer todos los criterios, sin embargo, la aplicación de los mismos es de gran utilidad para distinguir la información de alta a la de baja calidad.

En el mundo de los negocios, por otro lado, el avance tecnológico y de Internet ha tenido un gran impacto permitiendo la aparición de diversas aplicaciones Web generando mayor competitividad, la cual, debe ser asegurada a través de la calidad de los datos en las aplicaciones utilizadas [MAC05].

A través del estudio de diferentes trabajos orientados a características de calidad en los documentos Web, podemos rescatar la siguiente información: precisión, completitud y puntualidad, las cuales se fijan como las más estudiadas, destacándose en la mayoría de los trabajos. En menor medida encontramos las características consistente, conciso, actualizado, interpretabilidad, relevancia, seguridad, accesibilidad, cantidad apropiada de datos, disponibilidad, credibilidad, objetividad, reputación, fiabilidad de la fuente y trazabilidad (rastreable). En menor medida, detectamos como características de calidad en los documentos

Web la confidencialidad, la conveniencia, el grado de duplicados y de granularidad y la flexibilidad entre otros. Cabe destacar que varias de las características nombradas anteriormente se incluyen en las normas internacionales de calidad de software, sin embargo, no todas se consideran en el mismo nivel en este tipo de normas, por ejemplo disponibilidad y comprensibilidad. En resumen, hay una serie de convenciones universalmente aceptadas de las características deseables de la información, entre las cuales encontramos la *objetividad* en primer lugar, la *integridad* y la *utilidad*.

Por otro lado, y como aporte complementario, podemos considerar las valoraciones más frecuentes de un sitio Web entre las cuales encontramos el *mantenimiento y la actualización*, especificando cuando fue revisado por última vez, si el contenido es estático o dinámico; *si existen enlaces a otros sitios*, si estan *actualizados*, si la información presentada es *simple y confiable*, si son de *utilidad y mostrados de forma clara y visible*, si son *relevantes para el tema en cuestión*; *diseño y estructura de las páginas del sitio*, refiriéndose a como esta organizada y estructurada la información, permitiendo un acceso fácil y lógico; *estabilidad*, haciendo alusión a si a lo largo del tiempo se mantiene la URL, *acceso al sitio*, facilidad de conexión, descarga rápida, *aporte gráfico*, si las imágenes o audios incluídos en el sitio hacen un aporte importante y apropiado, si hace ameno el contenido presentado y por último, *requisito adicional de software o hardware*, haciendo hincapié en que los sitios de calidad deberían poder ser accesibles para cualquier usuario de Internet, independientemente de los programas o de la computadora utilizada.

Existen problemas que afectan directamente la calidad de la información y los datos en la Web, como consecuencia de la naturaleza de la Web, como lo son información desactualizada, problemas de diseño que hacen difícil al usuario el acceso a la información, publicaciones de información inconsistente, vínculos obsoletos, etc [Epp02]. En el campo del comercio electrónico, si los datos no son de calidad, es posible que la empresa tenga grandes pérdidas económicas y de imagen [Lim00, Hai03]. También, el dinamismo en la Web [Ami03], el uso e interacción de diferentes fuentes externas, los servicios en tiempo real y en particular, los cambios que sufren las fuentes y datos constantemente [Per02, Ger04] pueden afectar directamente la calidad de la información.

Otra de las situaciones típicas encontradas en la Web es la integración de datos estructurados con datos no estructurados [Fin] además de la conexión de datos diferentes [Zhu, AM04, BP04, Ger04]. En ambos casos, el desafío mayor, se encuentra en conectar información con diferentes niveles de calidad y entregarlo al usuario listo para su uso. Bajo este mismo contexto, es importante el enfoque desde la perspectiva desde los usuarios y la ayuda para entender los datos y su calidad, el uso de un lenguaje común para facilitar la comunicación entre las personas, sistemas, etc [AM04, Cap04].

Los métodos de evaluación y mejora de la calidad de la información en una aplicación Web son muy variados. Por ejemplo, en el trabajo de [Kat99], pro-

ponen un instrumento de evaluación para sitios Web individuales, a través de un formulario con 41 preguntas divididas en grupos, [Zhu] establecen un conjunto de criterios para la evaluación y selección de recursos Web como fuente de datos externa, considerando también, criterios de evaluación agrupados en tres categorías (estabilidad de la fuente, calidad de datos y aplicaciones específicas). Estos métodos son utilizados por organizaciones y managers de tecnologías de la información para capturar el estado de la información.

Queda claro que la Web debe adquirir mecanismos que hagan frente a estos cambios para lograr una evolución drástica hacia la calidad de los datos utilizados en la misma. Surge al mismo tiempo, la necesidad de definir un marco general de calidad para los datos en la Web que incluyan todas las características, inclusive aquellas con diferentes perspectivas del usuario.

Las dimensiones nombradas de la calidad de la información pueden ser agrupadas en cuatro categorías: intrínseca, contextual, representacional y accesibilidad.

La calidad intrínseca de la información, hace referencia al valor objetivo de la información independientemente de su forma, difusión, diseño o público al que se lo dirige.

- *Rigor científico:* Es importante que dicha información este respaldada por evidencia científica, en métodos científicos basados en cada una de las disciplinas.

- *Integridad:* la información no debe ser parcial ni sesgada, presentada en su totalidad siempre y cuando no sea otro el objetivo (resumen).

- *Objetividad:* un aspecto que no es fácil de percibir, la credibilidad de la información se asocia a la confianza que tenemos en el responsable del contenido en función de su autoridad. La carencia de esta característica puede llevar a percepciones erróneas o desinformación. Cabe destacar que la objetividad depende del autor de la misma y no de la percepción del usuario.

- *Precisión:* este aspecto se relaciona con la exactitud de la información y con la profundidad con las que se aborda un tema. Esta dimensión depende de la intención y las pretensiones del recurso y a los usuarios a la que va dirigido.

La calidad contextual de la información hace alusión al contexto en el que se accede a la información y a la adecuación del usuario. Encontramos aquí los siguientes aspectos:

- *Relevancia:* implica cuanto se adecua la información a las necesidades del usuario, dimensión subjetiva ya que depende del tipo de usuario.

- *Valor añadido:* estos elementos facilitan el uso de la información para una mejor asimilacion de la misma, mejorando la calidad.
- *Actualidad de la información:* exceptuando la información con valor histórico, esta caraterística determina calidad sobre todo en información científica o noticias.
- *Cantidad de información aportada:* a mayor información aportada, mejor será, siempre teniendo en cuenta los límites del sistema en el que se cargan los datos.
- *Utilidad:* esta característica tiene tanto un aspecto subjetivo, ya que mide cuan útil es la información y esto depende del tipo de usuario, mientras que por otro lado, la finalidad y el perfil del usuario al que se dirige aporta un aspecto objetivo.
- *Adecuación al usuario:* este aspecto depende puramente del usuario, pero es importante aclarar a quién va dirigida la información y adecuarla a ese perfil.

Calidad representacional de la información. Se relaciona con la forma en que se representa la información, así como también todos los aspectos técnicos referidos a su estructura. En esta categoría encontramos la claridad, concisión, compatibilidad, diseño, flexibilidad, homogeneidad de los datos y tipos de formatos.

Calidad del acceso a la información. En esta categoría se engloban los aspectos relativos al cómo se accede a la información, como lo son *tiempo de espera*, *navegación* y *seguridad* [Mol].

Actualmente, resulta de vital importancia que dispongamos de criterios claros y funcionales para la evaluación y selección debido a la facilidad de publicar y la cantidad de documentos en la Web, para que el usuario pueda discernir la veracidad, credibilidad, fiabilidad y la calidad de las informaciones de este medio. Esto puede verse claramente en entornos profesionales, como el académico y el científico que requieren de recursos de información mas rigurosos y pertinentes, o el mundo empresarial y comercial cuyos clientes exigen una información veraz, organizada y de calidad [Mol].

En el ambiente Web, aunque no todos persiguen los mismos objetivos y criterios, muchos son los profesionales que se dedican a evaluar los contenidos electrónicos. De forma general, se distinguen los grupos de profesionales, expertos, sociedades científicas, universidades, documentalistas y agencias de evaluación. Cada uno de estos grupos aporta un valor añadido a los recursos electrónicos de portales y directorios.

Respecto a los criterios de evaluación, la calidad de la información electrónica, puede ser evaluada desde diferentes perspectivas, aunque la más común es aquella que se centra en la satisfacción del usuario. Debemos considerar que siempre antes de comenzar a analizar, se debe identificar la tipología del documento ya que

existe una gran variedad de gamas de recursos como lo son páginas personales, Web comercial, etc. Los criterios e indicadores variarán en función de las características de los recursos así como también del nivel de profundidad que desee emplear el evaluador. La naturaleza hipertextual de la Web exige considerar dos niveles de evaluación sobre los que hay que aplicar criterios e indicadores:

- *Micronavegación:* se refiere a todos los aspectos relacionados con la navegación interna entre los propios contenidos del sitio Web.
- *Macronavegación:* relacionada con los enlaces del sitio Web hacia el exterior y la visibilidad del mismo en todo el entorno de la Internet.

A continuación se presentan algunos criterios y parámetros aplicados a cada uno de estos niveles.

- *Autoría:* indispensable para distinguir la credibilidad de la fuente de información. El responsable de los contenidos debe ser claramente identificado mediante algun indicador, como por ejemplo breve descripción del curriculum del responsable, una adscripción del autor a la organización a la que pertenece, algun logotipo que represente a la institución, etc.
- *Actualización:* se relaciona a la actualidad de los contenidos del sitio Web. Entre los indicadores más comunes encontramos la indicación explícita de la fecha de creación del sitio Web, la indicación explícita de la fecha de actualización de los contenidos, la correcta existencia de vínculos, etc.
- *Contenido:* Este criterio engloba varios requerimientos propios de los contenidos proporcionados en un sitio Web. Entre los más destacados encontramos:
  - *Cobertura:* Este indicador valora el nivel de los contenidos tratados en el sitio Web.
  - *Exactitud, precisión y rigor:* la exactitud de los contenidos incluidos en un sitio Web deben poder ser verificables, como por ejemplo, desde el punto de vista científico, apoyarse en citas bibliográficas, etc.
  - *Pertinencia:* Se relaciona con la validez y utilidad de los contenidos incluidos en el sitio Web, valorando tanto los propósitos declarados por el creador de los contenidos como el interés que posea la información para el usuario.
  - *Objetividad:* se intenta identificar en la información, presencia o ausencia de cualquier sesgo ideológico, político o comercial.

- *Accesibilidad:* se refiere a todas las características que hacen que el sitio Web sea accesible, sin dificultades para los diferentes usuarios. Entre lo más destacado nos encontramos con: diseño compatible con diferentes navegadores o diferentes resoluciones de pantalla, existencia de versiones alternativas de visualización para los sitios Web, posibilidad de imprimir, existencia de ayuda al usuario, etc.

- *Funcionalidad:* este criterio valora la efectivdad del sitio Web a la hora de utilizarlo y consultarlo. Entre los indicadores básicos que se deben tener en cuenta para valorar este criterio encontramos una estructura lógica de los contenidos, pertinencia y adecuación de títulos utilizados, la existencia de un mapa Web y de un sistema de búsqueda de contenidos propios del sitio Web.

- *Navegabilidad:* este criterio hace alusión a la facilidad con que el usuario se puede desplazar en las páginas de un sitio Web, aspecto relacionado con el conjunto de recursos y estrategias de navegación para conseguir un resultado óptimo a la hora de encontrar la información. Los indicadores que se rescatan de este criterio son presencia de un menú de contenidos siempre visible asi como tambien presencia de botones de navegación que permitan al usuario recorrer el sitio Web de manera lógica.

- *Diseño:* aquí se valora que el recurso, el documento Web, sea un espacio agradable a la vista y fácil de leer por el usuario Dentro de este apartado se valoran varias cuestiones relacionadas con el aspecto físico o la ergonomía del sitio Web que contribuyen a hacer del recurso digital un espacio agradable a la vista y fácil de leer por el usuario. Los indicadores que aportarían información engloban un diseño Web elegante, funcional y atractivo, adecuada combinación de colores, formas e imágenes que faciliten la lectura de los contenidos, Tipografía adecuada, homogeneidad de estilo y formato en todas las páginas del sitio Web. [Mol]

Como elemento final, en el cuadro 2.3 detallamos los tipos de problemas de la calidad de la información junto con sus posibles causas y acciones tomadas.

2.3. Calidad de la información en Wikipedia

Gracias a su gran escala, continua evolución y de colaboración abierta, Wikipedia se ha convertido en un recurso muy popular en nuestros días proporcionando información valiosa a los usuarios de Internet, además de plantear nuevos e importantes retos en diferentes áreas de la organización de la información, incluyendo la calidad de la información.

La definición de calidad, como se mencionó anteriormente, depende del contexto, por lo que no existe ningún modelo fijo de garantía de *Información de*

Calidad que se pueda aplicar a todos los sistemas. En Wikipedia el contexto está bien definido bajo el género enciclopedia. Sin embargo, la calidad de un artículo enciclopédico definido en esta Web difiere actualmente al definido por Crawford [Cra01b] orientado a la enciclopedia escrita en papel, para la cual se determinan siete criterios generales: ámbito, formato, unicidad, autoridad, exactitud, vigencia y accesibilidad.

La calidad de la información en Wikipedia se formalizó a través del *artículo destacado*. Estos son considerados como los mejores artículos que Wikipedia puede ofrecer.

Dichos artículos deben cumplir con[1]:

1. Estar bien escrito, completo, bien investigado, neutral y estable.
2. Proporcionar una sección concisa, una estructura apropiada y citas apropiadas.
3. Contar con imágenes y otros medios visuales.
4. Tener una longitud razonable y nivel de detalle.

El Cuadro 2.2 muestra la clasificación de los diferentes niveles de calidad que puede tener un artículo en Wikipedia[2].

Los principales componentes del marco de aseguramiento de la calidad de Wikipedia son coordinación de trabajo y artefactos de soporte, roles y procesos. Este marco asume al menos 3 tipos de procesos:

- Los que evaluan la calidad de un artículo y actúan directamente sobre este modificándolo, eliminándolo o cambiando su estatus.
- Los que evaluan la performance de los editores de Wikipedia y seleccionan agentes que aseguran la calidad (administradores, bots, etc.).
- Los que construyen y mantienen los artefactos de la coordinación de trabajo de Wikipedia.

Estos procesos nombrados pueden afectar la calidad de los artículos de Wikipedia directa o indirectamente, modificando su contenido así como también el flujo de trabajo. Identifiquemos a continuación elementos existentes en Wikipedia que podrían afectar la calidad de la información en Wikipedia.

Con una comunidad "hambrienta" a contribuir, Wikipedia permite a los usuarios elegir en que artículos y procesos contribuir y que roles tomar. Entre los roles detectados encontramos:

[1]Artículo destacado. https://en.wikipedia.org/?title=Wikipedia:Featured_articles.

[2]Niveles de calidad en un artículo.
http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment.

Categoría	**Descripción**	**Artículos**
FA	*Featured Article, artículo destacado.* Un artículo marcado como la mayor obra de Wikipedia. Distinguido por estándares profesionales de escritura, presentación y fuentes.	4321
A	*A-Class, Clase A.* El artículo está bien organizado y completo, habiendo sido revisado por evaluadores imparciales de WikiProject u otros.	1054
GA	*Good Article, Buen Artículo.* Un buen artículo es un artículo satisfactorio que ha alcanzado los buenos criterios, pero no ha encontrado el criterio de artículo destacado.	16753
B	*B-Class, Clase B.* El artículo está completo en su mayoría pero requiere cierto trabajo futuro para alcanzar los estándares de un buen artículo.	83523
C	*C-Class, Clase C.* El artículo es importante, pero aún contiene una gran cantidad de material irrelevante. El artículo debería tener referencias a fuentes fiables. Puede tener problemas significativos o requerir limpieza sustancial.	132444
Start	*Start, Inicio.* Un artículo que está desarrollado, pero que se encuentra bastante incompleto y podría requerir fuentes confiables.	904086
Stub	*Stub, Esqueleto.* Una descripción muy básica del tema.	2206104
FL	*Featured List, Lista de características.* Cubre un tema que conduce a la lista de formatos. Son conocidas como las mejores listas en Wikipedia.	1790
List	*List, Lista.* Cumple con los criterios de una lista independiente, que es un artículo que contiene principalmente una lista, que por lo general consta de enlaces a artículos en un área temática particular.	104209
		3454284

Cuadro 2.2: Esquema de clasificación de la calidad de los artículos que componen la Wikipedia en Inglés, junto con una descripción de los criterios respectivos y el correspondiente número de artículos que posee dicha categoría.

- *Editores:* son aquellos agentes que contribuyen directamente en los artículos.
- *Agentes de aseguramiento de la calidad:* son agentes que afectan la calidad del artículo, como lo son lo que revierten instancias de vandalismo, mejoras de la calidad del artículo a traves de ediciones menores, etc.
- *Agentes maliciosos:* son aquellos que degradan la calidad de un artículo a propósito.
- *Agentes ambientales:* son aquellos que modifican la calidad de un artículo por cambios producidos en el mundo real. Aunque la mayoría de las veces estos agentes degradan la calidad de los artículos, en algunos casos pueden mejorarlos alineando el estado del mundo real con la información contenida en un artículo.

Por otra parte, se puede afirmar que Wikipedia no provee un sistema con usuarios con diferentes niveles de permisos, sin embargo, podemos encontrar tres grupos de cuentas:

- *Usuarios registrados:* identificados por el nombre con el que ingresan a Wikipedia
- *Usuarios anónimos:* identificados por el protocolo IP.
- *Usuarios administradores y burócractas:* son usuarios registrados que poseen diferentes privilegios dentro de Wikipedia para editar los artículos.

Tanto los usuarios anónimos como los usuarios registrados, pueden asumir diferentes roles en cualquier momento, por ejemplo, los usuarios registrados podrían ser editores, agentes de calidad de la información o agentes maliciosos. En el caso de los administradores, aunque es probable que degraden un artículo, generalmente se espera que mejoren los artículos, cumpliendo un rol de editores o agentes de la calidad de la información. Aunque como dijimos anteriormente, Wikipedia permite que los usuarios seleccionen ellos mismos los roles y procesos, también promueve mecanismos para detectar buenos de malos editores y de esta forma moderar el trabajo sobre la información. Aquellos editores que esten dispuestos a mejorar la calidad de la información son promovidos a administradores dependiendo de su performance, conocimiento de Wikipedia, etc. Los administradores, poseen privilegios que les permiten, entre otras cosas, proteger, borrar, editar y restaurar artículos. Por otro lado, los burócratas, son administradores como se indicó anteriormente, pero tienen todos los permisos de administradores, pudiendo realizar cualquier tarea dentro de Wikipedia.

Vale destacar que el modelo de aseguramiento de la calidad de Wikipedia tiene dos mecanismos para controlar y hacer cumplir la calidad de los artículos:

- Asignación de un estatus “Artículo destacado", los cuales poseen un nivel alto deseado de calidad.
- Procesos de eliminación del artículo, que fuerzan a un mínimo nivel de calidad.

Es importante destacar que el número de artículos destacados dentro de Wikipedia es muy pequeño, rondando el 1 % del total de los artículos de Wikipedia en Inglés. Dichos artículos destacados, son aquellos que corresponden a los mejores artículos que Wikipedia puede ofrecer, según lo determinado por los Wikipedians, editores de Wikipedia. Los candidatos a artículos destacados son revisados para determinar precisión, neutralidad, integridad y estilo, de acuerdo a los criterios de los artículos destacados. Actualmente se cuenta con 4308 artículos destacados de aproximadamente 4.500.000 en Inglés. Esto nos indica que se posee 0.1 % de ellos, un valor muy bajo. El artículo destacado es representado en el sitio Web a través del icono de una estrella de bronce en la esquina superior derecha. Además, si el artículo, también es destacado en otros idiomas, una estrella aparecerá luego del vínculo del lenguaje en la lista posicionada del lado izquierdo de la página Web.

Los artículos pueden ser nominados como candidatos a artículo destacado por individuos o un grupo. Luego de nominados, los artículos candidatos son sometidos a una revisión para comprobar si cumplen con los criterios de artículos destacados. Según los registros de los artículos destacados, el proceso se inició en abril del 2002, aunque durante ese tiempo ningún artículo se sometió a un proceso de revisión por pares[3], ni se hacia referencia a ningún criterio de evaluación, solo eran llamados artículos con "buena prosa".

Una definición mas general utilizada es “aptitud para el uso” [Jur92], mientras que una definición mas operacional, incluye dimensiones de evaluación de calidad del contexto.

La primera guía para evaluar la calidad de la información fue escrita por la elocuencia del usuario en el año 2004. Esta guía incluía criterios como completitud, ajuste a los hechos y bien escrito. Esta guía fue sometida, luego, a múltiples cambios, actualmente incluyendo 8 criterios de evaluación: completo, preciso y verificable con referencias, estable sin cambios frecuentes, bien escrito, no controversial con lenguaje neutro, cumplimiento de normas estandares de Wikipedia, tener imágenes correctas con su copyright y por último, que tenga una longitud apropiada.

Aunque algunas de las dimensiones enumeradas en el modelo de Wikipedia como lo son estable, no controversial y verificable, no se enumeran en el marco de Crawford [Cra01a], podemos deducir que se dan por sentadas. La figura 2.1 mapea las dimensiones de Crawford [Cra01a] con el modelo de Wikipedia nombrado y el marco posterior creado por Stvilia [Gas01].

[3]Revisión por pares. http://es.wikipedia.org/wiki/Revisión_por_pares

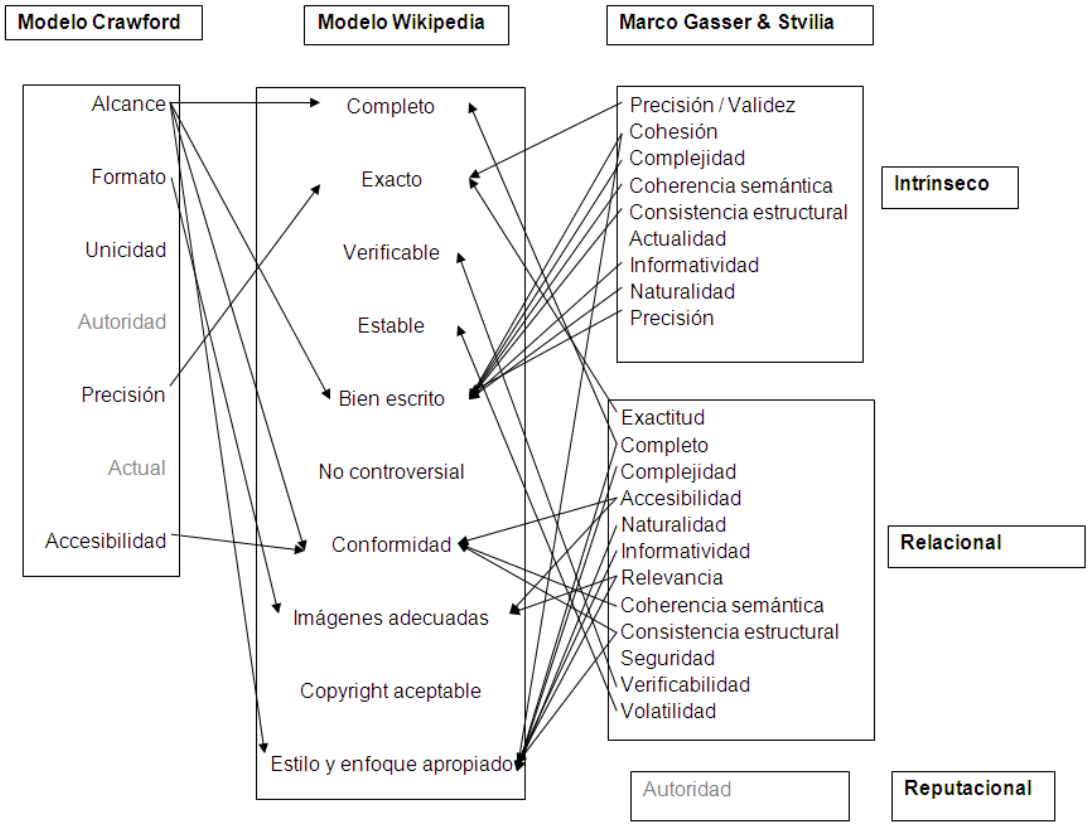

Figura 2.1: Mapeo de las dimensiones de Crawford contra el modelo de Wikipedia y el marco creado por Stvilia

Por otro lado, en Wikipedia, también se puede nominar un artículo destacado para que sea removido de ese status. Entre los años 2004 / 2005, aproximadamente 120 artículos fueron nominados para ser removidos, y solo a 80 de ellos se les quitó la categoría de artículo destacado.

Wikipedia, en particular, mantiene sus propios criterios y políticas de calidad. A continuación se explican y clarifican las tres políticas de contenido básico, el *punto de vista neutral*, *verificabilidad* y *páginas de investigacion no original*, aspectos y prácticas de Wikipedia descriptos por consenso comunitario. Estas políticas determinan el tipo y la calidad de los artículos de Wikipedia, trabajando en armonía de manera conjunta y no de forma aislada.

- *Punto de vista neutral:* Todos los artículos de la Wikipedia y otros contenidos enciclopédicos deben ser escritos desde un punto de vista neutral, significativo, sin ningún sesgo. Explicado de otra manera, la neutralidad implica un análisis de fuentes confiables que intenta transmitir esa información de manera justa y proporcional. Es, como Wikipedia afirma, "no

Problema	**Causa**	**Acción Tomada**
Accesibilidad	Organización Pobre. Políticas internas de Wikipedia. Barrera del lenguaje. Copyright.	Reorganizar, duplicar, remover, traducir, unir.
Exactitud	Errores de tipeo. Cambios en el mundo real. Bajo dominio del idioma. Ilegible por software.	Ajustar, corregir, cambiar, actualizar, verificar, explicar, clarificar el contexto, remover.
Autoridad	Falta de fuentes de soporte. Sesgo conocido de la fuente. Generalización infundada.	Reformular, calificar, añadir, sustituir, eliminar,
Cohesión	Pérdida de foco.	Restringir, mover.
Completitud	Múltiples perspectivas. Falta de detalles. Cobertura desequilibrada de diferentes perspectivas.	Añadir, especificar, eliminar ambigüedad, incluir, equilibrar, calificar, aclarar, integrar, exponer.
Complejidad	Baja Legibilidad. Lenguaje complejo.	Reemplazar, reescribir, simplificar, mover, resumir.
Consistencia	No cumplir normas sugeridas. Diferencias en la semántica del lenguage. Contradicción de hechos, cultura o norma social.	Reorganizar, voto, conformar, revertir, elegir la forma mas usada.
Informatividad	Redundancia de contenido.	Remover, revisar, mover.
Naturalidad	Lenguaje oscuro, texto no fluye bien.	Editar, mejorar, reescribir.
Relevancia	Agregar contenido no relevante.	Revertir, separar, eliminar.
Verificabilidad	Falta de referencias y acceso a las fuentes.	Agregar, remover, citar.
Volatilidad	Inestabilidad causada por guerra de ediciones o vandalismo.	Evitar, proteger.

Cuadro 2.3: Categorías y dimensiones para la calidad de la información y datos de calidad.

negociable", es decir, que no puede ser reemplazada por otras políticas o directrices. Obviamente, cada editor posee su propio punto de vista, sin embargo, debe intentar no promover un punto de vista particular propio. Teniendo en cuenta esto, el *punto de vista neutral* no implica la exclusión de ciertos puntos de vista, sino la inclusión de todos los puntos de vista verificables. A continuación se enumeran los principios que permiten alcanzar el nivel de neutralidad apropiado:

- *Evitar establecer opiniones como hechos:* generalmente los artículos de Wikipedia incluyen opiniones significativas acercas de sujetos y/o hechos. Estas deberían ser solamente mostradas en el texto de fuentes particulares. Por ejemplo, un artículo no debe indicar que el genocidio es una mala acción, pero puede afirmar que si lo es para un autor particular.
- *Evitar indicar afirmaciones discutidas como hechos:* si en diferentes fuentes se encuentran afirmaciones contradictorias de un tema, éste debe ser tratado como opinión y no como hecho.
- *Evitar afirmar hechos como opiniones:* las afirmaciones no controversiales realizadas por fuentes confiables deben expresarse como tales en Wikipedia, excepto que se trate un tema en el que hay desacuerdo en la información.
- *Preferir un idioma sin prejuicios:* el punto de vista neutral no simpatiza ni menosprecia el tema o lo que afirman las fuentes confiables.
- *Indicar la importancia de puntos de vista opuestos:* se debe asegurar que los diferentes puntos de vista sobre un tema reflejan niveles de apoyo debidos y no se le da una importancia indebida a un punto de vista particular. Por ejemplo, decir que "de acuerdo con Simón Wiesenthal, el Holocausto fue un programa de exterminio de los judíos en Alemania, pero David Irving rebate este análisis"sería dar aparente paridad entre la opinión de mayoría calificada y de una minoría.

Como regla general, no se debe retirar información de Wikipedia por el solo hecho que parece tener un sesgo, en lugar de eso, se debe intentar reescribir o citar material para producir una perspectiva mas neutral. A continuación se enumeran algunos problemas comunes y posibles soluciones:

- *Nombrado:* a veces el nombre de un tema puede mostrar parcialidad. Si un nombre se utiliza a menudo en fuentes confiables y es reconocido por los lectores, puede ser utilizado aunque puedan considerarlo sesgado. Por ejemplo los títulos: “Masacre de Boston” y “Jack el destripador".

  Los títulos descriptivos deben ser redactados en forma neutral para no sugerir un punto de vista a favor o contra un tema, por ejemplo, "Las

criticas a X", puede ser reescrito como "puntos de vista sociales sobre X".

- *Estructura del artículo:* la neutralidad debe ser protegida a través de atención adicional en la estructura interna de un artículo y así evitar problemas como puntos de vista diversos. El sesgo en diferentes secciones o subsecciones, puede dar lugar a una estructura no enciclopédica, mostrando dialogos de "ida y vuelta", favoreciendo indebidamente a un punto de vista.
- *Peso debido o indebido:* la neutralidad exige que cada artículo represente todos los puntos de vista significativos de las fuentes confiables, asignándoles la debida importancia y así evitar dar importancia indebida a opiniones minoritarias. Estas deberían incluirse en un apartado "ver también". Por ejemplo, el artículo acerca de la tierra, no menciona apoyo actual al concepto de tierra plana, si lo hiciera una clara minoría, sería darle indebida importancia a la misma.

  El peso indebido también puede darse por no incluir profundidad en el detalle, por la cantidad de texto incluida y yuxtaposición en las sentencias.

  En resumen, siempre se debe ser claro en que partes del texto se describen puntos de vistas minoritarios, además de que la opinión de la mayoría debe ser explicado con suficiente detalle que el lector pueda comprender la diferencia entre ambos puntos.
- *Aspectos de balanceo:* Un artículo no debe dar importancia indebida a cualquier aspecto del tema tratado, es decir, debe tratar cada aspecto con un "peso" adecuado. Por ejemplo, analizar hechos aislados de un tema puede ser verificable, pero aún desproporcionada en relación con su importancia general para el tema del artículo.
- *Atribuir igual validez puede crear un falso equilibrio:* la política de Wikipedia no implica exponer cada punto de vista minoritario, aún cuando es importante tener todos los puntos de vista. Por ejemplo, que la tierra es plana, que los caballeros templarios poseían el Santo Grial, etc. Dichas historias, son especulativas, pero en la actualidad no aceptadas, las teorías deben ser legitimadas académicamente.
- *Una buena investigación:* buena e imparcial investigación, basándose en las mejores y más reputadas fuentes autorizadas disponibles ayuda a evitar desacuerdos sobre el punto de vista neutral.
- *Balance:* la neutralidad asigna peso a los diferentes puntos de vista en proporción a su importancia. Pero, cuando fuentes de buena reputación se contradicen entre sí, se describen ambos enfoques para trabajar y mantener el equilibrio, basándose en fuentes secundarias y terciarias.

- *Tono imparcial:* como Wikipedia incluye disputas,p ero no participa de las mismas, los conflictos requieren la presentación de puntos de vista con un tono imparcial.

  Los artículos neutrales se escriben con un tono objetivo, preciso y proporcionado de todas las posiciones del mismo, intentando no citar a los participantes involucrados, pero si resumir y presentar los argumentos imparcialmente.

- *Describiendo opiniones estéticas:* diferentes tipos de artículos, como por ejemplo los que involucran música o arte, tienen tendencia a ser efusivos, lo cual no corresponde a una enciclopedia. Es difícil ponerse de acuerdo sobre quien es el mejor soprano del mundo, pero, es favorable decir como cierto soprano fue recibido por el público en general. Las críticas públicas y académicas verificables proporcionan un contexto útil para el arte.

- *Palabras para revisar:* no existen las palabras prohibidas en Wikipedia, aunque ciertas expresiones deben utilizarse con mucho cuidado para no introducir sesgos. Por ejemplo, si decimos "John dijo que no había comido el pastel"podemos implicar que el había comido el pastel, no en el caso de decir "John dijo que no comió el pastel".

- *Sesgo en las fuentes:* un argumento en una disputa acerca de fuentes confiables implica una fuente sesgada. El sesgo en el argumento de fuentes es una forma de presentar un punto de vista neutral, excluyendo las fuentes que se disputan entre sí. El punto de vista neutral debe lograrse mediante el balance basado en el peso de la opinión de las fuentes confiables y no excluyendo las fuentes que no conforman el punto de vista del escritor.

- *Verificabilidad:* en Wikipedia, este aspecto, significa que el material debe por ser contrastado con una fuente confiable, publicada. Wikipedia no publica investigaciones originales, sino que su contenido, se determina por la información publicada previamente en lugar de las creencias o experiencias de los editores. Incluso, si el editor sabe que es verdad, debe ser verificable [Wik]. Demostrar la verificabilidad recae en el editor que agrega el material, y esta se cumple proporcionando una cita a una fuente confiable (que también debe estar publicada) que apoya la contribución. Cabe destacar que todo material que necesita una fuente, pero no tiene una, puede ser eliminado, aunque como paso intermedio, se puede dejar una etiqueta que indique que falta la cita. Las citas se deben realizar de forma clara y precisa (especificando la página, sección o divisiones), además de apoyar claramente el material que se presenta en el artículo.
  En Wikipedia, la cita posee tres significados, los cuales, si faltasen pueden afectar la fiabilidad:

- El tipo de trabajo (Documento, artículo, libro).
- El creador / escritor.
- El editor de la obra.

Las mejores fuentes tienen una estructura profesional en el lugar para comprobar o analizar los hechos, asuntos legales, pruebas y argumentos.

Otras fuentes confiables para Wikipedia incluyen:

- Los libros de texto de nivel universitario
- Libros publicados por editoriales respetados
- Revistas
- Diarios de amplia circulación

Por otro lado, las fuentes que generalmente no son confiables son:

- Fuentes cuestionables
- Fuentes auto publicadas
- Fuentes de auto publicadas o cuestionables como fuentes en sí mismas
- Wikipedia misma o sitios que reflejan contenidos de Wikipedia

- *Investigación no original:* Como se dijo en el punto anterior, el material publicado en Wikipedia debe ser atribuído a una fuente confiable, publicada. Wikipedia no publica ningún pensamiento original, todo debe estar basado o atribuible a fuentes confiables.

  La frase “investigación original"(OR) se utiliza en Wikipedia para referirse al material que contiene hechos e ideas para los cuales no existen fuentes confiables publicadas. Para demostrar que no se esta agregando OR, se debe publicar fuentes confiables que estan directamente relacionadas con el tema del artículo presentado.

  Por ejemplo: la afirmación “la capital de Francia es París” no necesita ninguna fuente, porque nadie se opondrá a ella y sabemos que existen fuentes que la comprueben. La declaración se puede atribuir, aunque no sea atribuido. Más alla de la necesidad de atribuir contenidos a fuentes confiables, estos no se deben plagiar o violar los derechos de autor, los artículos deben ser escritos con palabras propias manteniendo el significado original del material en cuestión. A modo de regla general, cuanto mayor cantidad de participantes existan en la comprobación del hecho , mas confiable es la fuente.

  En general, las fuentes más fiables son las siguientes:

- Revistas
- Libros publicados por editoriales universitarias
- Los libros de texto de nivel universitario
- Revistas, diarios y libros publicados por editoriales respetados
- Diarios de amplia circulación

Los artículos en Wikipedia, deben estar basados en fuentes secundarias fiables, publicadas y en menor grado, en fuentes terciarias y primarias.

- *Fuentes primarias:* corresponden a materiales originales, cercanos a un evento y generalmente están escritos por personas directamente involucradas en él. Las fuentes primarias pueden o no ser fuentes independientes o de terceros, por ejemplo, una historia de un accidente de tránsito escrito por un testigo, es una fuente primaria de información sobre el accidente. Los documentos históricos tales como los diarios son fuentes primarias.

  Como política general, las fuentes primarias que se publicaron de forma fiable en Wikipedia pueden usarse, pero con cuidado, ya que es fácil hacer mal uso de ellas. Además, cualquier interpretación de fuente primaria, requiere una fuente secundaria fiable. La fuente primaria, solo se puede utilizar para hacer declaraciones directas, descripciones de hechos que pueden ser verificadas en las fuentes primarias, etc. Por ejemplo, en un artículo donde se describe una novela, toda interpretación debe tener una fuente secundaria.

- *Fuentes secundarias:* ofrecen pensamientos propios de un autor basado en fuentes primarias, muy cercanos al evento. Contiene la interpretación de un autor, el análisis o evaluación de los hechos, pruebas, conceptos e ideas de fuentes primarias. Cabe destacar que las fuentes secundarias no son necesariamente fuentes independientes o de terceros, se basan en fuentes primarias para el material, por ejemplo, un libro de un historiador militar de la Segunda Guerra Mundial, sería una fuente secundaria, pero si incluye detalles de las propias experiencias de guerra del autor, seria fuente primaria.

  La política de Wikipedia se basa en mantener fuentes secundarias fiables y los reclamos analíticos o evaluativos de los artículos, solo puede hacerse si han sido publicados por una fuente secundaria fiable.

- *Fuentes terciarias:* son publicaciones que resumen las fuentes primarias y secundarias, sobre todo cuando estas fuentes se contradicen. Por ejemplo, Wikipedia es una fuente terciaria y muchos libros de texto de nivel universitario también son considerados como tales porque resumen múltiples fuentes secundarias.

Estas políticas determinan el tipo y la calidad del material que es aceptable en los artículos de Wikipedia. Estas, se complementan entre sí y no deben interpretarse de forma aislada, así como también, los principios en que se basan estas declaraciones políticas no son reemplazadas por otras políticas o directrices.

La *investigación no original* tiene sus orígenes en la política *punto de vista neutral* y en el problema de lidiar con teorías de peso e indebidas. Esta última política intenta proveer un marco en el que los editores con diversos puntos de vista puedan colaborar en la creación de una enciclopedia. El principio base utilizado, sostiene que es difícil que las personas se pongan de acuerdo respecto a que es verdad, pero si es posible que las personas se pongan de acuerdo en relación a lo que ellos y otros creen que es verdad, por lo cual Wikipedia no utiliza la "verdad" como criterio de inclusión. En lugar de eso, se utiliza la política de imparcialidad. Años posteriores se detectó que la política *punto de vista neutral* se estaba violando y que era necesario complementarla con una nueva política que se llamó *verificabilidad* con el fin de asegurar la exactitud de los artículos a través de la introducción de citas en los mismos.

Capítulo 3

Conclusiones

3.1. Contribuciones

Índice de cuadros

Bibliografía

[AD09] Anthony D., Smith S., W. T. *Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia.* Rationality & Society 21, 2009.

[AFE10] Anthony Fader, S. S., and Etzioni, O. *Identifying Relations for Open Information Extraction.* 2010.

[AM04] Angeles, P., and MacKinnon, L. *Detection and Resolution of Data Inconsistences, and Data Integration using Data Quality Criteria.* QUATIC'2004, 2004.

[Ami03] Amirijoo, M. *Specification and Management of QoS in Imprecise Real-Time Databases.* Proceeding of the Seventh International Database Engineering and Applications Symposium, 2003.

[And13] Anderka, M. *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia.* 2013.

[AS10] Anderka, M., and Stein, B. *A Breakdown of Quality Flaws in Wikipedia.* 2010.

[Blu08] Blumenstock, J. E. *Size matters: word count as a measure of quality on Wikipedia.* ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367673, 2008.

[BM07] Banko M., Cafarella M. J., S. S. B. S. y. E. O. *Open information extraction from the Web.* In the Proceedings of the 20th International Joint Conference on Articial Intelligence, 2007.

[Bou01] Bouzeghoub, M. y Kedad, Z. *Quality in Data Warehousing. Information and Database Quality.* M. Piattini, C. Calero y M. Genero, Kluwer Academic Publishers, 2001.

[BP04] Bouzeghoub, M., and Peralta, V. *A Framework for Analysis of data Freshness.* Proc. IQIS2004, 2004.

[BSG05] Besiki Stvilia, Michael B. Twidale, L. C. S., and Gasser, L. *Assessing information quality of a community-based encyclopedia.* MIT, 2005.

[BSG08] B. Stvilia, M. Twidale, L. C. S., and Gasser., L. *Information quality work organization in wikipedia.* 2008.

[Cap02] Cappiello, C. *A Model of Data Currency in Multi-Channel Financial Architectures.* 2002.

[Cap04] Cappiello, C. *Data quality assessment from the user's perspective.* Proc. IQIS2004, 2004.

[Cra01a] Crawford, H. *Encyclopedias.* 3ra Edición. Englewood, CO: Libraries Unlimited., 2001.

[Cra01b] Crawford., H. *Encyclopedias. In Richard E. Bopp and Linda C. Smith, editors, Reference and information services: an introduction.* 2001.

[DB03] Davis, G.F., Y. M., and Baker, W. *The small world of the American corporate elite 1982-2001.* Strategic Organization, Vol. 1 No. 3, 2003.

[Epp02] Eppler, M. y Muenzenmayer, P. *Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Aplication Methodology.* 2002.

[Etz12] Etzioni, O., B. M. S. S. W. D. *Open information extraction from the web.* 2012.

[Fel98] Fellbaum, C., e. *WordNet: an electronic lexical database.* MIT Press, 1998.

[Fin] Finkelstein, C. y Aiken, P. *XML and Corporate Portals.* http://www.wilshireconferences.com/xml/paper/xml-portals.htm.

[Fug02] Fugini, M. *Data Quality in Cooperative Web Information Systems.* 2002.

[FW10] Fei Wu, D. S. W. *Open Information Extraction using Wikipedia.* 2010.

[Gas01] Gasser, L., . S. B. *A new framework for information quality.* Champaign: University of Illinois at Urbana-Champaign., 2001.

[Ger04] Gertz, M. *Report on the Dagstuhl Seminar "Data Quality on the Web".* SIGMOD Record, vol. 33, Nº 1, 2004.

[Hai03] Haider, A. y Koronios, A. *Authenticity of Information in Cyberspace: IQ in the Internet, Web, and e-Business.* 2003.

[JG99] Juran, J. M., and Godfrey, A. B. *Juran's quality handbook.* McGraw-Hill London., 1999.

[Joh98] Johnson, R., W. D. *Applied multivariate statistical analysis (4th ed.).* Upper Saddle River, NJ: Prentice-Hall., 1998.

[Juf09] Juffinger, A., G. M. L. E. *Blog credibility ranking by exploiting verified content.* In: Proceedings of the Workshop on Information Credibility on the Web (WICOW) in conjunction with WWW'2009, New York, NY, USA, ACM (2009) 51-58, 2009.

[Jur92] Juran, J. *Juran on quality by design.* 1992.

[Kat99] Katerattanakul, P. y Siau, K. *Measuring Information Quality of Web Sites: Development of an Instrument.* Proceeding of the 20th International Conference on Information System., 1999.

[KJ93] Kim J., M. D. *Acquisition of semantic patterns for information extraction from corpora.* Proceedings of Ninth IEEE Conference on Artificial Intelligence for Applications, 1993.

[KS02] Kakihara, M., and Sørensen, C. *Exploring knowledge emergence: from chaos to organizational knowledge.* Journal of Global Information Technology Management, Vol. 5 No. 3, 2002.

[Lee02] Lee, Y. *AIMQ: a methodology for information quality assessment. Information and Management.* Elsevier Science, 2002.

[Lex12] Lex, E., V. M. E. M. F. E. C. L. H. C. S. B. G. M. *Measuring the quality of web content using factual information.* In 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12). ACM., 2012.

[LGD09] Loris Gaio, Matthijs den Besten, A. R., and Dalle., J.-M. *Wikibugs: using template messages in open content collections.* ACM. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641330., 2009.

[Lih04] Lih, A. *Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource.* Journalism and Media Studies Centre University of Hong Kong, 2004.

[Lim00] Lim, O. y Chin, K. *An Investigaction of Factors Associated With Customer Satisfaction In Australian Internet Bookshop Websites.* 3rd Western Australian Workshop on Information Systems Research, Autralia., 2000.

[LJ11] Liu J., R. S. *Who does what: Collaboration patterns in the wikipedia and their impact on article quality.* ACM Transactions on Management Information Systems 2, 2011.

[LP13] Lian Pohn, Edgardo Ferretti, M. E. *Evaluación de la Calidad de la Información de Wikipedia en Español.* http://sedici.unlp.edu.ar/bitstream/handle/10915/27332 /Documento_completo.pdf?sequence=1, 2013.

[LS10] Lipka, N., and Stein, B. *Identifying featured articles in Wikipedia:writing style matters.* ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772847., 2010.

[MAB12] M. Anderka, B. S., and Busse, M. *On the Evolution of Quality Flaws and the Effectiveness of Cleanup Tags in the English Wikipedia.* 2012.

[MAC05] M. Angélica Caro, Coral Calero, I. C. M. P. *Data Quality in Web Applications: A State of the Art.* 2005.

[Mag10] Magdy, A., W. N. *Web-based statistical fact checking of textual documents.* In: Proceedings of the 2nd international workshop on Search and mining user-generated contents. SMUC '10, New York, NY, USA, ACM (2010) 103-110, 2010.

[MAL11] Maik Anderka, B. S., and Lipka., N. *Towards automatic quality assurance in Wikipedia.* ACM. ISBN 978-1-4503-0637-9. doi: 10.1145/1963192.1963196., 2011.

[MAL12] M. Anderka, B. S., and Lipka, N. *Predicting Quality Flaws in Usergenerated Content: The Case of Wikipedia.* In 35th Annual SIGIR Conference. ACM, 2012.

[Mar03] Marchetti, C. *Enabling Data Quality Notification in Cooperative Information Systems through a Web-service based Architecture.* 2003.

[MBD11] Michael Bendersky, W. B. C., and Diao., Y. *Quality-biased ranking of web documents.* ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935849., 2011.

[MI12] Myshkin Ingawale, Amitava Dutta, R. R. P. S. *Network analysis of user generated content quality in Wikipedia.* 2012.

[Mol] Molina, M. P. *Calidad y evaluacion de los contenidos electrónicos.* http://www.mariapinto.es/e-coms/eva_con_elec.htm.

[OFR12] Oliver Ferschke, I. G., and Rittberger, M. *FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia.* 2012.

[Per02] Pernici, B. y Scannapieco, M. *Data Quality in Web Information Systems.* Proceeding of the 21st International Conference on Conceptual Modeling, 2002.

[PT10] Pirkola, A., and Talvensaari., T. *A topic-specific web search system focusing on quality pages.* Springer. ISBN 3-642-15463-8. doi: 10.1007/978-3-642-15464-5_64., 2010.

[R.97] R., H. *Evaluating Internet Research Sources.* http://www.sccu.edu/faculty/R_Harris/evalu8it.htm, 1997.

[SEMZ09] Stuart E. Madnick, Richard Y. Wang, Y. W. L., and Zhu., H. *Overview and framework for data and information quality research.* Journal of data and information quality, 1(1):1–22, 2009. ISSN 1936-1955. doi:10.1145/1515693.1516680., 2009.

[Sta00] Stacey, R. *The emergence of knowledge in organizations", Emergence: Complexity and Organization.* Vol. 2 No. 4, 2000.

[Str97] Strong, D. *Data Quality in Context.* Communications of the ACM. Vol. 40. Nº 5., 1997.

[Suc07] Suchanek, F.M., K. G. W. G. *Yago: A core of semantic knowledge.* In: Proceedings of the 16th International Conference on World Wide Web. WWW '07, New York, NY, USA, ACM (2007) 697-706, 2007.

[US05] Uzzi, B., and Spiro, J. *Collaboration and creativity: the small world problem.* American Journal of Sociology, Vol. 111 No. 2, 2005.

[WD07] Wilkinson D., H. B. *Cooperation and quality in Wikipedia.* In 3rd international symposium on wikis and open collaboration, 2007.

[WF94] Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, New York, NY., 1994.

[Wik] Wikipedia. *http://en.wikipedia.org/wiki/Wikipedia: Verifiability,_not_truth.*

[WS96] Wang, R. Y., and Strong, D. M. *Beyond accuracy: what data quality means to data consumers.* Journal of management information systems, 12(4):5–33, 1996. ISSN 0742-1222., 1996.

[ZC05] Zhou, Y., and Croft., W. B. *Document quality models for Web ad hoc retrieval.* ACM. ISBN 1-59593-140-6. doi: 10.1145/1099554.1099652., 2005.

[ZG00] Zhu, X., and Gauch., S. *Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web.* ACM. ISBN 1-58113-226-3. doi: 10.1145/345508.345602., 2000.

[Zhu] Zhu, Y. y Buchmann, A. *Evaluating and Selecting Web Sources as external Information Resources of a Data Warehouse.*