

Mejorando la Navegación en Wikipedia con Etiquetas Sociales

Wikimania 2009

Arkaitz Zubiaga

NLP & IR Group @ UNED

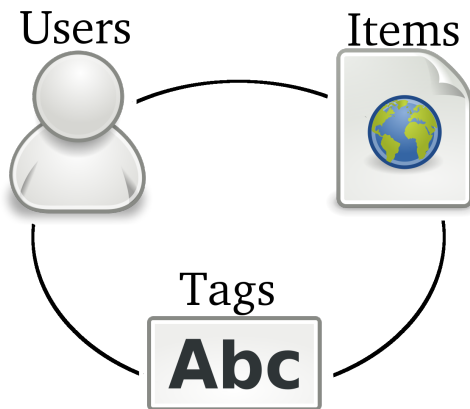
28 de agosto de 2009

<http://blog.zubiaga.org>

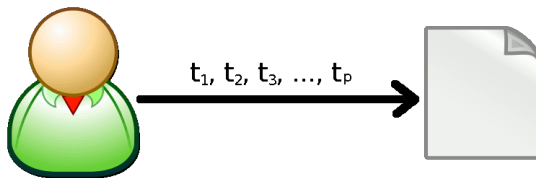
Índice

- 1 **Introducción**
- 2 Navegando por Wikipedia
- 3 Ventajas de las Etiquetas
- 4 Generación del Conjunto de Datos
- 5 Resultados
- 6 Conclusiones

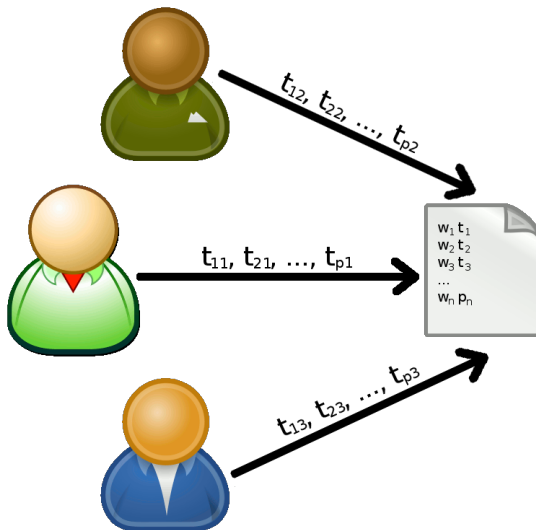
Un Sistema de Etiquetas



Etiquetado Simple



Etiquetado Colaborativo



Nube de Etiquetas

Tag Cloud: Popular

KEY: [blue tags](#) are tags you have in common with everyone else.

Sort: [Alphabetically](#) | [By size](#)

.net 2008 3d advertising ajax and animation api apple architecture art article articles artist audio **blog** blogging blogs book **books** browser **business** car cms code collaboration comics community computer converter cooking cool **css** **culture** data database **design** desktop **development** diy documentation download downloads drupal ebooks economics **education** electronics email entertainment environment fashion fic film finance firefox **flash** flex flickr **food** forum **free** freeware fun funny gallery game **games** geek **google** government graphics green guide hardware health history home hosting house **howto** html humor icons illustration images imported information **inspiration** interactive interesting **internet** iphone japan java **javascript** jobs jquery kids **language** **learning** library linux list lists literature mac magazine management maps marketing **math** media microsoft **mobile** money movie movies mp3 **MUSIC** network networking **news** online **opensource** osx people phone photo **photography** photos photoshop php plugin podcast **politics** portfolio privacy productivity **programming** psychology python radio rails realestate recipe recipes **reference** religion **research** resources reviews rss ruby rubyonrails school science **search** security seo shop **shopping** social socialnetworking **software** statistics streaming teaching tech **technology** tips todo tool **tools** toread travel **tutorial** tutorials tv twitter typography ubuntu usability **video** videos vim visualization web **web2.0** **webdesign** webdev **wiki** wikipedia windows wishlist wordpress work writing youtube

Índice

- 1 Introducción
- 2 Navegando por Wikipedia**
- 3 Ventajas de las Etiquetas
- 4 Generación del Conjunto de Datos
- 5 Resultados
- 6 Conclusiones

Navegando por Wikipedia

- Categorías
- Enlaces entre artículos
- Motor de búsqueda

Navegando por Wikipedia: Categorías

Computer programming

From Wikipedia, the free encyclopedia

Categories: [Software development process](#) | [Computer programming](#)

- Pro: Organización del contenido
- Con: Limitado a la taxonomía y su categorización.
 - Un artículo simplemente puede pertenecer o no a una categoría.

Navegando por Wikipedia: Enlaces entre artículos

Computer programming

From Wikipedia, the free encyclopedia

"Programming" redirects here. For other uses, see [Programming \(disambiguation\)](#).

Computer programming (often shortened to **programming** or **coding**) is the process of writing, testing, debugging/troubleshooting, and maintaining the [source code](#) of [computer programs](#). This source code is written in a [programming language](#). The code may be a modification of an existing source or something completely new. The purpose of programming is to



- Pro: Acceso a artículos relacionados
- Con: Sujeto a disponibilidad de enlaces

Navegando por Wikipedia: Motor de búsqueda

Search results

From Wikipedia, the free encyclopedia

Content pages [Multimedia](#) [Help and Project pages](#) [Everything](#) [Advanced](#)

There is a page named "[Programming](#)" on this wiki

[Computer programming](#) (redirect from [Programming](#))

Computer **programming** (often shortened to **programming** or coding) is the process of writing, testing, debugging/troubleshooting, and ...

20 KB (2555 words) - 16:35, 7 August 2009

[Programming](#) language

A **programming** language is a machine-readable artificial language designed to express computations that can be performed by a machine, ...

46 KB (5869 words) - 05:41, 6 August 2009

- Pro: Búsqueda de artículos por palabras clave
- Con: Sujeto a disponibilidad del término en contenido/categorías

Motivación

¿Se podría mejorar el acceso a los artículos mediante etiquetas?

Índice

- 1 Introducción
- 2 Navegando por Wikipedia
- 3 Ventajas de las Etiquetas**
- 4 Generación del Conjunto de Datos
- 5 Resultados
- 6 Conclusiones

Ventajas de las Etiquetas

Las etiquetas...

- ...son simples
- ...se basan en un vocabulario abierto
- ...se añaden de forma agregada
- ...permiten a los usuarios hacer diferentes definiciones

¿Cómo Podemos Navegar mediante Etiquetas?

Tres formas de navegar sobre etiquetas¹:

- Navegación pivotada (Pivot-browsing)
- Popularidad
- Filtrado

¹Tagging: People-Powered Metadata for the Social Web, by Gene Smith

Navegación por Etiquetas: Pivot-browsing

- Ventaja: Navegación cruzada sobre etiquetas/temas relacionados.

Tag:Programming

From Wikipedia, the free encyclopedia

- [List of algorithms \(1045\)](#)

[algorithms](#) [programming](#) [reference](#) [algorithm](#) [wikipedia](#) [math](#) [software](#) [code](#) [list](#)
[cs](#) [development](#) [search](#) [wiki](#) [coding](#) [theory](#) [useful](#) [computerscience](#) [computer](#)
[mathematics](#) [data](#) [science](#) [resource](#) [algoritmos](#) [design](#) [research](#) [computers](#)
[compsci](#) [documentation](#) [cool](#) [algo](#)

- [Ajax \(427\)](#)

[ajax](#) [programming](#) [javascript](#) [wikipedia](#) [web](#) [xml](#) [webdesign](#) [web2.0](#) [development](#)
[reference](#) [webdev](#) [wiki](#) [software](#) [design](#) [html](#) [technology](#) [css](#) [xmlhttprequest](#)
[internet](#) [article](#) [dhtml](#) [tutorial](#) [java](#) [tech](#) [definition](#) [code](#) [xhtml](#) [tools](#) [howto](#)
[architecture](#)

Navegación por Etiquetas: Popularidad

- Ventaja: ¿Qué han considerado los usuarios interesante/relevante para una etiqueta?

Tag:Programming

From Wikipedia, the free encyclopedia

- List of algorithms (1045)
algorithms programming reference algorithm wikipedia math software code list cs development search wiki coding theory useful computerscience computer mathematics data science resource algorithms design research computers compsci documentation cool algo
- Ajax (427)
ajax programming javascript wikipedia web xml webdesign web2.0 development reference webdev wiki software design html technology css xmlhttprequest internet article dhtml tutorial java tech definition code xhtml tools howto architecture
- Agile software development (355)
agile development software programming process wikipedia reference management methodology projectmanagement design productivity project softwaredevelopment work wiki xp business collaboration web scrum software-development code dev coding methodologies project_management documentation gtd article
- Design pattern (computer science) (302)
patterns programming design designpatterns development pattern software reference wikipedia wiki architecture java code design_patterns oop designpattern coding design-pattern webdev design-patterns documentation design_pattern article computer software_engineering oo software-engineering c++ software.engineering cs
- Representational State Transfer (293)
rest webservises web programming architecture http xml reference wikipedia web2.0 api development webdev design software restful wiki webservice ajax soa webdesign patterns article services framework soap dev documentation protocol network

Navegación por Etiquetas: Filtrado

- **Benefit:** Retrieving documents containing a tag(s), but not another. Hence, users can retrieve articles related to a topic, and excluding a subtopic.

Tag:Programming (excluding projectmanagement)

From Wikipedia, the free encyclopedia

Articles

- **List of algorithms** (1045)
 - algorithms programming reference algorithm wikipedia math software code list cs development search wiki coding theory useful computerscience computer mathematics data science resource
 - algorithms design research computers compsci documentation cool algo
- **Ajax** (427)
 - ajax programming javascript wikipedia web xml webdesign web2.0 development reference webdev wiki software design html technology css xmlhttprequest internet article dhtml tutorial java tech definition code xmlhttp tools honte architecture
- ~~**Agile software development** (355)~~
 - agile development software programming process wikipedia reference management methodology projectmanagement design productivity project softwaredevelopment work wiki ip business collaboration web scrum software-development code dev coding methodologies project_management documentation gtd article
- **Design pattern (computer science)** (302)
 - patterns programming design designpatterns development pattern software reference wikipedia wiki architecture java code design_patterns oop designpattern coding design-pattern webdev design-patterns documentation design_pattern article computer software_engineering oo software-engineering c++ software.engineering cs
- **Representational State Transfer** (293)
 - rest webservice web programming architecture http xml reference wikipedia web2.0 api development webdev design software restful wiki webservice ajax soa webdesign patterns article services framework soap dev documentation protocol network
- **Comparison of web application frameworks** (232)
 - framework comparison web programming frameworks webdev web2.0 development php reference ajax java application architecture webdesign wikipedia design wiki webdevelopment software ruby python rails list rubyonrails cms library resources django code
- **Comet (programming)** (213)
 - comet ajax programming javascript push web http web2.0 development architecture webdev wikipedia server streaming browser article technology wiki webdevelopment framework serverpush messaging client application polling java apache webdesign event asynchronous
- ~~**Anti pattern** (200)~~
 - programming patterns design wikipedia development reference pattern management software architecture designpatterns antipattern code antipatterns anti-patterns anti-pattern java coding article technology theory humor wiki anti funny business projectmanagement design-patterns designpattern design-patterns
- **Bloom filter** (194)
 - programming algorithm algorithms filter bloom hash wikipedia math search data performance reference bloomfilter datastructures structure cs development probability database set science datastructure research bloom-filter hashing computer cache article computer-science probabilistic

¿Cómo Pueden las Etiquetas Mejorar las Taxonomías? (1)

● Definición

- Las categorías deben crearse antes de utilizarse. Esto no ocurre con las etiquetas, las cuales se crean la primera vez que se utilizan
- Cada usuario define un conjunto de etiquetas, generando como resultado un listado global de etiquetas con pesos. Esto permite que una etiqueta sea más relevante que otra, lo cual no es posible con las categorías, ya que es necesario un acuerdo entre usuarios en su definición.
- Las etiquetas pueden manejar sinonimia, hiperónimos, hipónimos,... a diferencia de las categorías.

¿Cómo Pueden las Etiquetas Mejorar las Taxonomías? (y 2)

- Navegación y búsqueda
 - Proveen las nuevas formas de navegación mencionadas anteriormente.
 - Proveen nuevos términos que describen los artículos.
 - Proveen una clasificación de abajo a arriba (bottom-up), en lugar de arriba a abajo (top-down).

Un Ejemplo que Combina Etiquetas y Categorías: Amazon


amazon.com Hello. Sign in to get [personalized recommendations](#). New customer? [Start here](#). FREE

Your Amazon.com | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

Shop All Departments Search All Departments

Books Advanced Search Browse Subjects New Releases Bestsellers The New York Times® Bestsellers Libros En Español

Click to LOOK INSIDE!



Tagging: People-powered Metadata for the Social Web (Voices That Matter) (Paperback)
 by [Gene Smith](#) (Author)
 ★★★★★ (5 customer reviews)

List Price: ~~\$39.99~~
 Price: **\$26.39** & this item ships for **FREE with Super Saver Shipping**. [Details](#)
 You Save: **\$13.60 (34%)**

In Stock.
 Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Wednesday, August 12? Order it in the next **12 hours and 13 minutes**, and choose **One-Day Shipping** at checkout. [Details](#)

34 new from \$21.99 **15 used** from \$15.50

Tags Customers Associate with This Product [\(What's this?\)](#)

Click on a tag to find related items, discussions, and people.

Check the boxes next to the tags you consider relevant or enter your own tags in the field below.

- | | | |
|---|---|---|
| <input type="checkbox"/> tagging (17) | <input type="checkbox"/> social software (6) | <input type="checkbox"/> social media (2) |
| <input type="checkbox"/> folksonomy (12) | <input type="checkbox"/> findability (5) | <input type="checkbox"/> bananas (1) |
| <input type="checkbox"/> metadata (12) | <input type="checkbox"/> ia (4) | <input type="checkbox"/> ong (1) |
| <input type="checkbox"/> information architecture (7) | <input type="checkbox"/> interaction design (2) | See all 26 tags... |

Popular in this category: [\(What's this?\)](#)

#60 in [Books](#) > [Nonfiction](#) > [Social Sciences](#) > [Communication](#) > [Technology & Society](#)


Evitando el Desorden de Etiquetas

- Permitiendo a los usuarios definir relaciones entre etiquetas, como lo hace Librarything²:



Tag info: science fiction

Includes: science fiction, science fiction, - science fiction, Ficcao Cientifica, Fiction (Science Fiction, Fiction (Science Fiction), Fiction-Science fiction, Fiction; Science Fiction, Ficção Cientifica, SF - Science Fiction, Sciecn Fiction, Science Fiction, Scirnce Fiction, "science fiction", ciencia ficcio, ciencia ficcion, ciencia ficción, ciencia-fiction, ciència ficció, fantascienza, fiction - science fiction, science fiction, sciene fiction, scienc fiction, science ficton, science fictino, science fictio, science fiction., science ficton, science+fiction, science-fiction, science.fiction, science_fiction, sciencefiction, sciences fiction, scienc fiction, sciene fiction, sf science-fiction, science fiction, sience fiction, sience-fiction (what?)

Tag and its aliases used 691,346 times by 19,941 users. 

²<http://www.librarything.com>

Índice

- 1 Introducción
- 2 Navegando por Wikipedia
- 3 Ventajas de las Etiquetas
- 4 Generación del Conjunto de Datos**
- 5 Resultados
- 6 Conclusiones

Generación del Conjunto de Datos

- Punto de inicio: conjunto de 2M+ artículos de la Wikipedia en inglés³.
- Obtención de datos:
 - Información de etiquetas para cada artículo desde Delicious⁴.
 - Contenido del artículo.
- Se eliminaron algunas etiquetas no relevantes para artículos de Wikipedia: *wikipedia*, *reference*, *wiki*.
- Se limitó a artículos anotados por al menos 10 usuarios.
- Resultado: 20,764 artículos de Wikipedia con sus etiquetas⁵.

³<http://en.wikipedia.org>

⁴<http://delicious.com>

⁵El conjunto de datos está disponible en <http://nlp.uned.es/social-tagging/>

Índice

- 1 Introducción
- 2 Navegando por Wikipedia
- 3 Ventajas de las Etiquetas
- 4 Generación del Conjunto de Datos
- 5 Resultados**
- 6 Conclusiones

Resultados: Nube de Etiquetas

Special:Tag Cloud

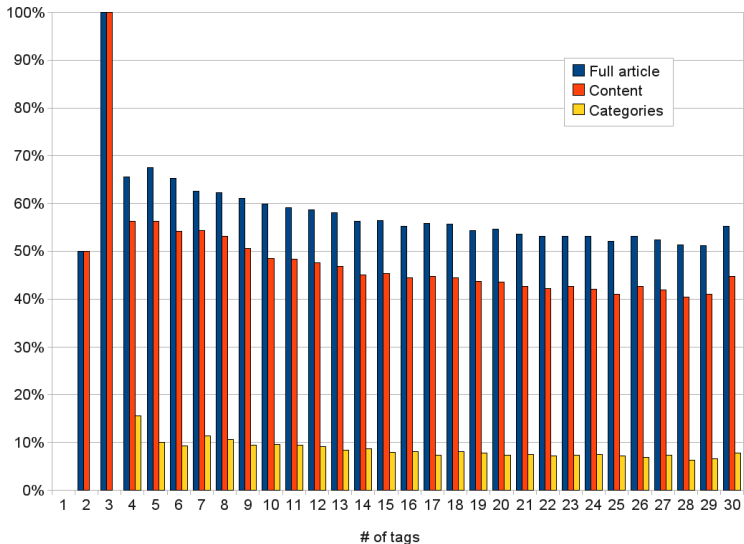
From Wikipedia, the free encyclopedia

agile ai ajax algorithm **algorithms** architecture art article articles artist audio
 biology book **books** brain **business** code collaboration community comparison
 computer cool cryptography **culture** data database definition **design** **development**
economics education encyclopedia energy english environment evolution film finance
 folksonomy food framework free fun funny future game **games** geek google graphics gtd
 hardware **health** **history** howto html humor ideas information inspiration
interesting internet japan java javascript knowledge **language** law learning lifehacks
 linguistics **linux** list literature logic **management** marketing **math**
mathematics media memory microsoft mind mobile money movies **music** mythology network
 networking **opensource** patterns people philosophy photography physics
politics productivity **programming** psychology religion research resources
science scifi search **security** semanticweb social socialnetworking sociology **software**
 space standards statistics tagging tech **technology** theory tools travel tv typography unix
 video visualization war **web** **web2.0** webdesign webdev weird windows words work writing
 xml

Resultados: Presencia de Etiquetas

	Encontrados	No encontrados	% Encontrado
En documento	251,139	206,569	54.86 %
En contenido	202,151	255,557	44.16 %
En categorías	35,237	422,471	7.70 %

Resultados: Presencia de Etiquetas



Resultados: Algunos Ejemplos

- Los usuarios han definido la etiqueta *programming* para el artículo *Framework*, el cual no está presente en el contenido.
- Lo mismo ocurre con la etiqueta *mathematics* para el artículo *Zipf's law*.
- Así como con la etiqueta *audio* para el artículo *List of Internet stations*.

Prototipo

Un prototipo preliminar de etiquetas aplicadas sobre Wikipedia está disponible en:

<http://taggedwiki.zubiaga.org>

Índice

- 1 Introducción
- 2 Navegando por Wikipedia
- 3 Ventajas de las Etiquetas
- 4 Generación del Conjunto de Datos
- 5 Resultados
- 6 Conclusiones**

Conclusiones

- Un sistema de etiquetado social proveería nuevas formas de navegación a Wikipedia:
 - Navegación pivotada (Pivot-browsing)
 - Popularidad
 - Filtrado
- Clasificación de abajo a arriba (bottom-up) en lugar de arriba a abajo (top-down).
- El experimento realizado con etiquetas sobre Wikipedia muestra resultados alentadores.
 - Provee nuevos términos no existentes.
 - Provee nuevas formas de navegación mediante etiquetas.
 - Ayuda a mejorar el motor de búsqueda.
 - Descubre artículos populares.

Muchas Gracias por su Atención

Achiu Arigato Danke Dhannvaad Dua Netjer en ek **Efcharisto**
Gracias Gràcies Gratia **Grazie** Guishepeli **Hvala** Kiitos
 Köszönöm **Mercé** Merci **Mila** esker Obrigado Shukran
 Shukriya Tack **Tak** Takk Tänan Tapadh leat Tesekkür ederim **Thank**
you Toda

<http://blog.zubiaga.org>