

Enhancing Navigation on Wikipedia with Social Tags

Wikimania 2009

Arkaitz Zubiaga

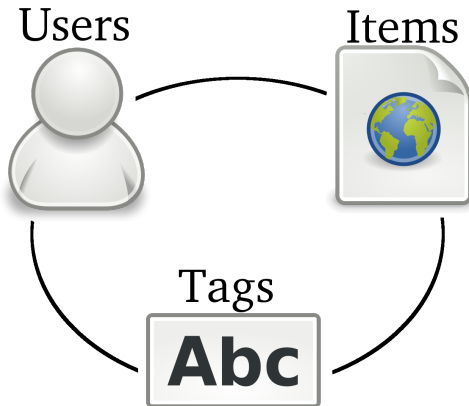
NLP & IR Group @ UNED

August 28th, 2009

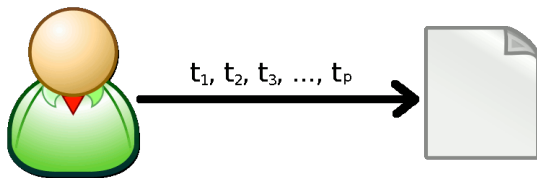
Index

- 1 Introduction
- 2 Navigating on Wikipedia
- 3 Benefits of Tagging
- 4 Dataset Generation
- 5 Results
- 6 Conclusions

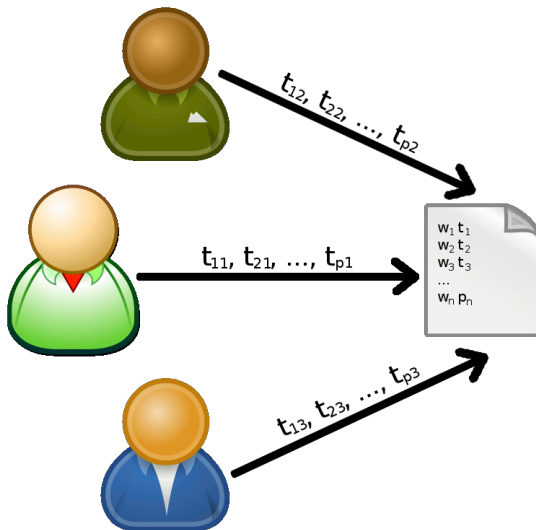
A Tagging System



Simple Tagging



Collaborative Tagging



Tag Cloud

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

Index

- 1 Introduction
- 2 Navigating on Wikipedia**
- 3 Benefits of Tagging
- 4 Dataset Generation
- 5 Results
- 6 Conclusions

Navigating on Wikipedia

- Categories
- Links in articles
- Search engine

Navigating on Wikipedia: Categories

Computer programming

From Wikipedia, the free encyclopedia

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

- Pro: Great for organizing content
- Con: Limited to the taxonomy and categorization.
 - An article is or is not in a category.

Navigating on Wikipedia: Links in articles

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: Access to related articles
- Con: Subject to link availability

Navigating on Wikipedia: Search engine

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 computation s that can be performed by a machine, ...

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

- Pro: Article search by keyword
- Con: Subject to term availability in content/categories

Motivation

Can article retrieval be improved by means of tagging?

Index

- 1 Introduction
- 2 Navigating on Wikipedia
- 3 Benefits of Tagging**
- 4 Dataset Generation
- 5 Results
- 6 Conclusions

Benefits of Tagging

Tags...

- ...are simple
- ...rely on an open vocabulary
- ...can be aggregated
- ...allow users to differ on the definition

How Can We Navigate Through Tags?

Three ways to navigate through tags¹:

- Pivot-browsing
- Popularity
- Filtering

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

Tag Navigation: Pivot-browsing

- Benefit: Switching to related tags/topics.

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](#)

Tag Navigation: Popularity

- Benefit: What has people considered interesting/relevant for a tag?

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
- 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

Tag Navigation: Filtering

- **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](#) (255)~~
 - ~~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 designpatterns 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

How Can Tagging Enhance Taxonomies? (1)

- Definition

- Categories must be created before using them, unlike tags, which are created the first time a user uses them
- Each user defines a set of tags, generating a global weighted set of tags as a result. This allows a tag to be more relevant than another, unlike with categories, for which an agreement among users is necessary.
- Tags can handle synonymy, hypernims, hyponims,... unlike categories.

How Can Tagging Enhance Taxonomies? (and 2)

- Navigation and search
 - Providing the aforementioned new ways to navigate.
 - Providing new terms to describe the articles.
 - Providing a bottom-up classification, instead of a top-down one.

An Example Combining Tags and Categories: 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 All Departments

Books | [Advanced Search](#) | [Browse Subjects](#) | [New Releases](#) | [Bestsellers](#) | [The New York Times® Bestsellers](#) | [Libros En Español](#)



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](#)


Avoiding Tag Mess

- Letting users to relate tags, as in 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, Science Fiction, Science Fiction, "science fiction", ciencia ficcio, ciencia ficcion, ciencia ficción, ciencia-fiction, ciència ficció, fantascienza, fiction - science fiction, sceince fiction, scicene fiction, scienc fiction, scienc ficiton, science fictio, science fiction., science ficton, science+fiction, science-fiction, science.fiction, science_fiction, sciencefiction, sciences fiction, scienc fiction, scinece fiction, sf science-fiction, sience fiction, sience fiction, sience-fiction (what?)

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

²<http://www.librarything.com>

Index

- 1 Introduction
- 2 Navigating on Wikipedia
- 3 Benefits of Tagging
- 4 Dataset Generation**
- 5 Results
- 6 Conclusions

Dataset Generation

- Starting point: a set of 2M+ English Wikipedia³ articles.
- Data retrieval:
 - Tag data for each article on Delicious⁴.
 - Article content.
- Removed not relevant tags for Wikipedia articles: *wikipedia*, *reference*, *wiki*.
- Filtered to articles annotated by at least 10 users.
- Result: 20,764 tagged Wikipedia articles⁵.

³<http://en.wikipedia.org>

⁴<http://delicious.com>

⁵The dataset is available for download at <http://nlp.uned.es/social-tagging/>

Index

- 1 Introduction
- 2 Navigating on Wikipedia
- 3 Benefits of Tagging
- 4 Dataset Generation
- 5 Results**
- 6 Conclusions

Results: Tag cloud

Special:Tag Cloud

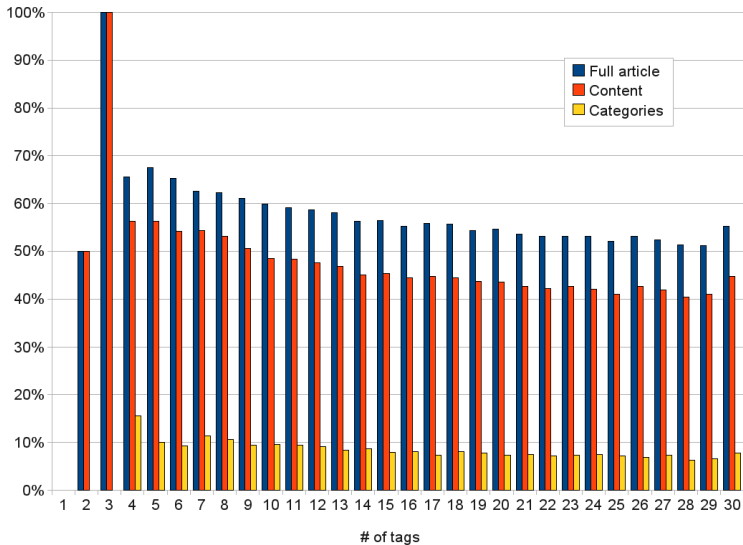
From Wikipedia, the free encyclopedia

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

Results: Presence of tags

	Found	Not Found	% Found
Document	251,139	206,569	54.86%
Content	202,151	255,557	44.16%
Categories	35,237	422,471	7.70%

Results: Presence of tags



Results: Some examples

- Users defined the tag *programming* for the article *Framework*, but that word is not present in content.
- The same occurs for the tag *mathematics* in the article *Zipf's law*.
- As well as for the tag *audio* in the article *List of Internet stations*.

Prototype

A preliminary prototype on applying tags to Wikipedia can be found at:

<http://taggedwiki.zubiaga.org>

Index

- 1 Introduction
- 2 Navigating on Wikipedia
- 3 Benefits of Tagging
- 4 Dataset Generation
- 5 Results
- 6 Conclusions**

Conclusions

- A social tagging system would provide new ways to navigate Wikipedia:
 - Pivot-browsing
 - Popularity
 - Filtering
- Bottom-up classification besides top-down classification.
- An experiment with tags over Wikipedia shows encouraging results.
 - Providing new unexisting terms.
 - Providing new ways to navigate through tags.
 - Helping to improve the search engine.
 - Discovering popular articles.

Thank You for Your Attention

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>