

PRESENTATION ON yaCy

PRESENTED BY:-
Nancy vaish
M.Tech (C.S.E) 1st Year



INTRODUCTION

- YaCy is a peer-to-peer search engine, meaning that there is no centralized authority or server where your information is stored.
- YaCy pronounced as (“Ya see”).
- Its core is a computer program written in Java distributed on several hundred computers so-called YaCy-peers. Each YaCy-peer independently crawls through the Internet, analyzes and indexes found web pages, and stores indexing results in a common database which is shared with other YaCy-peers using principles of P2P networks.
- YaCy is available on Windows, Mac and Linux.

COMPONENTS

YaCy search engine is based on four elements:

- Crawler
- Indexer
- Search and Administration interface
- Data Storage

yaCy SEARCH PAGE



**Web Search by the People,
for the People**

☒ Text ☐ Images [more options...](#)

As you can see, this is a pretty conventional search engine page. You can search using the provided search bar without any additional configuration, if you wish.

We will be exploring the administration interface though, because that provides us with a lot more flexibility. Click on the "Administration" link in the upper-left corner of the page:

You will be taken to the basic configuration page:

Peer Administration Console


Status Basic Configuration Accounts Network Configuration Download System Update
Performance Advanced Settings Local robots.txt Advanced Properties Thread Dump

Basic Configuration

Your YaCy Peer needs some basic information to operate properly

- ☒ Select a language for the interface:
☐ Deutsch ☐ Français ☐ 汉语/漢語 ☐ Русский ☐ Українська ☐ हिन्दी ☒ English
- ☒ Use Case: what do you want to do with YaCy:
☒ Community-based web search ☐ Search portal for your own web pages ☐ Intranet Indexing

Join and support the global network 'freeworld', search the web with an uncensored user-owned search network.

Your YaCy installation behaves independently from other peers and you define your own web index by starting your own web crawl. This can be used to search your own web pages or to define a topic-oriented search portal.

Create a search portal for your intranet or web pages or your (share) domain name or IP, or with an URL of the form file://<path> or smb. Files may also be shared with the YaCy server, assign a path here. This path can be accessed at <http://localhost:8090/repository>. U:
- ☐ Your peer name has not been customized; please set your own peer name
Peer Name:
- ☒ Your peer can be reached by other peers
Peer Port: ☐ with SSL (https enabled)
Configure your router for YaCy: ☒  Configuration was not successful. This may take a moment.

☒ Set Configuration

What you should do next:
Your Peer name is a default name; please set an individual peer name.

Crawl Sites to Contribute to the Global Index

You can now search using the indexes kept on your YaCy peers. The search results will become more and more accurate the more people participate in the system.

To start this process, click on the "Crawler / Harvester" link under the "Index Production" section.

Index Production	
	Crawler / Harvester
	Creation Monitor
	Index Administration
	Filter & Blacklists
	Content Semantic
	Target Analysis
	Process Scheduler

If you've attempted to search for something and did not get the results you were looking for, consider starting to index the pages on a site with your instance. It will make your search more accurate for yourself and your peers.

Type in the URL that you want to index in the "Start URL" section:

Site ☒ **Start URL (must start with**
http://
https://
ftp:// **smb://**
file://
Wikipedia, the free encyclopedia

☐ **Link-List of URL**
http://en.wikipedia.org/w/api.php?action=featuredfeed&feed=potd&feedformat=atom;http://en.wikipedia.org/wiki/Wikipedia:Featured_articles;http://en.wikipedia.org/wiki/Wikipedia:F130_Hercules_crash;http://en.wikipedia.org/wiki/Federal_popular_initiative_%22Against_mass15;http://en.wikipedia.org/wiki/Wikipedia:Picture_of_the_day/February_2014;http://en.wikipedia.title=Main_Page&oldid=593411281;http://en.wikipedia.org/w/index.php?title=Main_Page&oldidtitle=Main_Page&printable=yes;http://simple.wikipedia.org/wiki/http://ar.wikipedia.org/wiki/http://

☐ **Sitemap URL**

Path ☒ **load all files in domain**
☐ **load only files in a sub-path of given url**

Limitation ☐ **not more than** **documents**

Collection

Start ☒ **Start New Crawl**

You can limit the number of documents that your crawl will index. Click "Start New Crawl" when you are finished to begin crawling the selected site. Click on the "Creation Monitor" link on the left-hand side to see the progress of the indexing. You should see something like this:

Crawler

Queues

Queue	Size	Pause/Resume
Local Crawler	800	
Limit Crawler	0	
Remote Crawler	0	
No-Load Crawler	0	
Loader (200)	0	

pause reason: network switch to defaultsyacy.network.freeworld.uniz

Index Size

Database	Entries	Segments
Documents	98,703	12
solr_search_api		
Webgraph Edges	0	0
solr_search_api		
Citations (reverse link index)	28,326	1
RWIs (P2P Chunks)	63,064	2

Progress

Indicator	Level
Speed / PPM (Pages Per Minute)	6000 PPM 0.5 LF 20 MH set (min/max)
Crawler PPM	117
Postprocessing	idle
Progress	0.0 pending in collection: 3854 pending in webgraph: 0
Traffic (Crawler)	317.91 MB
Load	0.05

Running Crawls (2)

Name	Status
xxcd.com	Running Terminate
www.digitalocean.com	Running Terminate

Crawled Pages

Title	URL
xxcd: Five Years	http://xxcd.com/1088/
xxcd: Writing Styles	http://xxcd.com/1083/
xxcd: Physicists	http://xxcd.com/793/
xxcd: What xxcd Means	http://xxcd.com/207/
xxcd: Cold	http://xxcd.com/1321/
xxcd: Action Movies	http://xxcd.com/311/
xxcd: War	http://xxcd.com/769/
xxcd: Movie Seating	http://xxcd.com/173/
xxcd: Turtles	http://xxcd.com/889/
xxcd: Time	http://xxcd.com/1190/
xxcd: Asshole	http://xxcd.com/677/

Your server will crawl the URL specified at the rate of 2 requests per second until it has either run out of links chained together or reached the limit you set.

ADVANTAGES

- As there is no central server, the reliability is higher, because there's no single point of failure .
- Because the engine is not owned by a company, there is no centralized advertising.
- It is possible to achieve a high degree of privacy.
- On every search YaCy fetches the pages provided in search results and verifies that they still contain the keywords requested by the user. This ensures that the pages that no longer contain the requested keywords are not displayed to the user

DISADVANTAGES

- As there is no central server and the YaCy network is open to anyone, malicious peers are able to insert inaccurate or commercially biased search results.
- Result verification is done client-side on every search, which increases network traffic on the computer running YaCy and makes YaCy slower to display the search results than search engines such as Google.
- Missing IPv6 support.

CONCLUSION

If you need a great search engine for your site, YaCy provides that option as well. YaCy is very flexible and is an interesting solution to the problem of privacy concerns.

THANK YOU

