

Spinn3r architecture and data

Kevin Burton, Founder/CEO

spinn3r

What is Spinn3r?

- Licensed weblog, forum, and social media crawler
 - Save \$40k per month
- 300k posts per hour
- 21TB of content (1.2TB per month)
- 18 months of archives
- 3B documents
- +150Mb/s - 24/7

spinn3r

Theory of Operation

- Index content as quickly as possible
- Make compromises for latency and throughput
- No spam
- Discard no metadata

spinn3r

Hardware

- 40 mid-range (scale diagonally) Intel servers
- 22TB of raw storage ~60TB effective
- 200GB of in-memory data
- Three replicas
- Fault tolerant database
- Highly available

spinn3r

Live indexing

- Receive pings from social media sites
- Index content cyclically (30 minutes) for sites without pings
- Traditional crawlers must make sacrifices (crawl rate)
- Hybrid approach works well

spinn3r

Indexing Rates

- ~2-5M HTTP requests per hour
- 2-4k HTTP requests per second
 - RSS
 - Permalink URLs
 - New source discovery
 - Spam detection (90% of the ping stream)
 - Ping handling

spinn3r

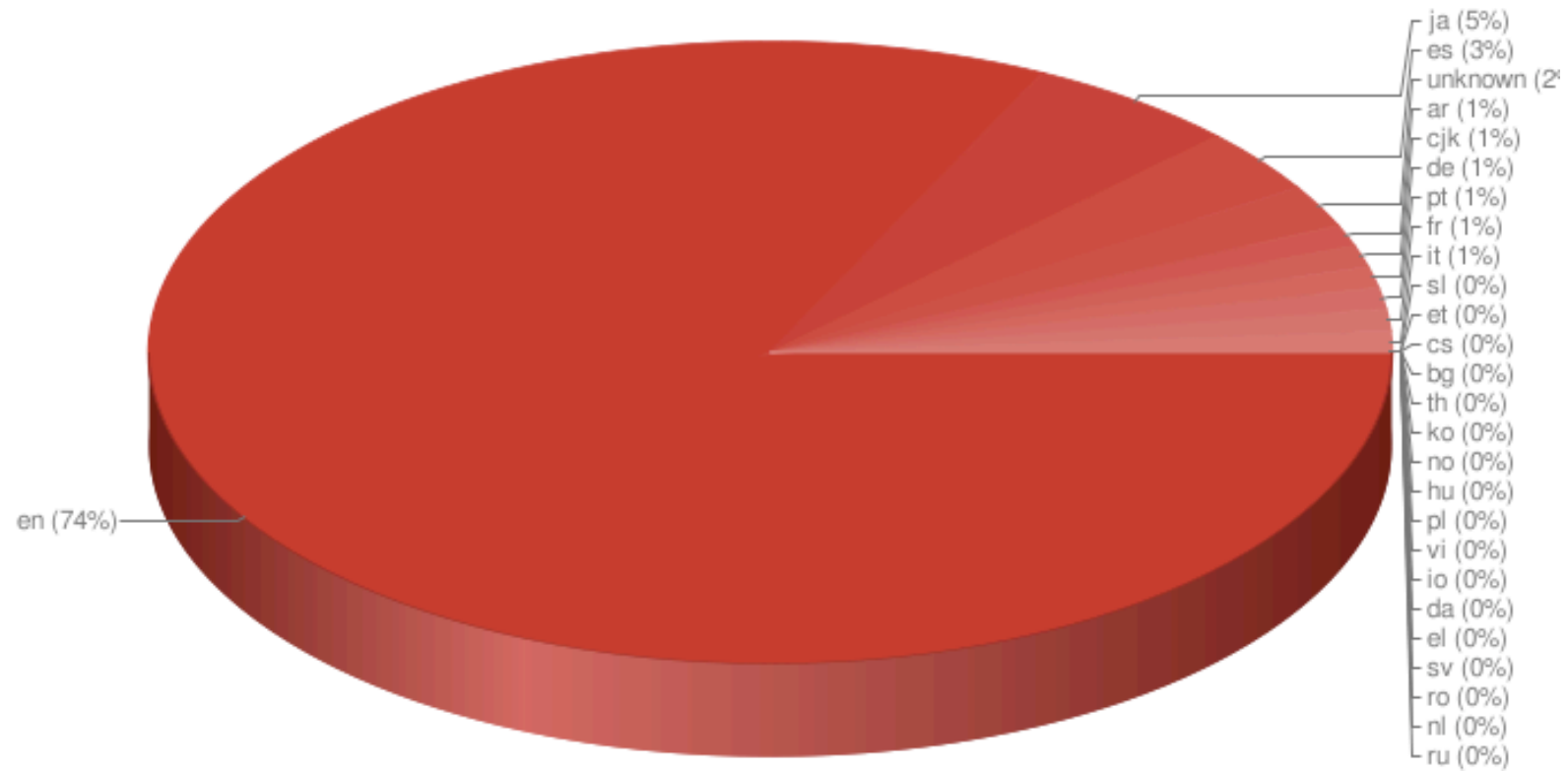
RSS and Atom

- Rich metadata
 - Accurate title
 - Tags
 - Publication time
 - Huge waste of bandwidth

Language classification

- Do not trust manually selected languages
- N-gram model
- Code page detection
- In production for more than three years

Permalink Language Breakdown (1 hour)



Fighting Spam

- Link analysis
- Text analysis
- Long tail content is the hardest

spinn3r

Spam Statistics

- 30% of our time is spent fighting spam
- 95% of pings are from spammers
- Primarily stolen content
- 10% malware
 - BAD when it happens

spinn3r

Smart Spammers

- Don't assume you can win
- Spammers are getting smarter
- Your elegant theory will be torn to shreds in practice
 - Pragmatism rules

Content Extraction

- High ranking sites disable full content in RSS/Atom feeds
 - Increases ad revenue
 - Reduced bandwidth cost
 - Probability that you will have summary content is directly proportional to your rank
- Full content is needed for search, sentiment analysis, link graph, etc.

spinn3r

Identify Full Content

- Strip all redundant HTML
- Only return content
- Result should be well formed XHTML including `` `` `<a>` elements

Ranking

- Time based rank
- Indegree
- Multiple stable ranking vectors
 - Language
 - Category
 - Time

spinn3r

Comments

- RSS/Atom feeds
- Template parsing
- Comment hosting

spinn3r

What's next

- More data for ICWSM in 2010
 - Comments
 - Content extract
 - Full HTML
 - 4TB
- Tighter duplicate content suppression
- New ranking
- Clustering

spinn3r