apt-xapian-index

Everything You Always Wanted to Index About Debian Packages, But were Afraid to Ask

Enrico Zini
enrico@debian.org

23 February 2008

Outline

Please help me with the notes

1. apt-get install gobby
2. Run gobby
3. Connect to the session at `192.168.42.217`, port `6522`, password `enrico`
4. Join document `notes.txt`

Outline

The problem

What I want to see happening

- Build smart interfaces to browse the large Debian archive.

The first problem I think needs solving:

- The only fast package index we have at the moment is `APT`
- The task of the `APT` index is to solve dependencies
- `APT` shouldn't be expanded (bloated) to do much more
- Solution: create another index to complement `APT`

What the new index should have

- Fast full text searches
- Fast tag searches
- Extensible, to accomodate new ideas for data to index

Outline

A tour of apt-xapian-index

The technology

- Sits in /var/lib/apt-xapian/index
- Based on Xapian
  - Indexes text as well as numbers and dates
  - Decent bindings in all sorts of languages
  - Stretchable and abusable by great lengths
- Self documented in /var/lib/apt-xapian-index/README

A tour of apt-xapian-index

Indexing

- Done by /usr/sbin/update-apt-xapian-index
- Can be run interactively
- Runs in a weekly cron job
- Packages can inject extra data by adding plugins in /usr/share/apt-xapian-index/plugins

A tour of apt-xapian-index

Searching

- You just need the plain Xapian API
- /var/lib/apt-xapian-index/README documents the index layout

Tools using it

- goplay (golearn, goadmin, ...)
- debtags.debian.net (just started)

Outline

This page is sneakily left blank to divert your attention elsewhere.

Getting more data into the system

My proposal

- One package per dataset to get
- Ship a copy of the dataset in the package, to use if everything fails
- A tool that can be run to fetch the data, or
- A plugin system to fetch the data using a single tool instead?
- Download new versions using a cron job
- Provide the data somewhere under /var
- Add an apt-xapian-index plugin to index it

For example: popcon, bts statistics, iterating.org

More indexing ideas

Debian specific stemming

- “lib*foo” becomes “library” and “foo”; “deb*foo” becomes “debian” and “foo”
- “cvsdelta”, “cvsgraph”, “gnomecatalog”, “gnomeradio”, “gnu*something” (but not “gnustep”), “kde*something”...
- More generally, how to index “Rindfleischetikett-tierungsüberwachungsaufgabenübertragungsgesetz”?
- How to provide the same stemming algorithm at query time?
- Compensate with improved descriptions?

More indexing ideas
What else to index?

- popcon
- bts statistics
- iterating.com
- more ideas?

i18n

How about searching translated descriptions?

- Xapian already supports stemming for many languages
- Is it useful, with such short descriptions?
- One index per language?
- How about disk space, and indexing time?

Index update
Can it be improved?

- Incremental updates
  - Need to track what's new after an *apt-get update*
  - Increases index size
- Suid update script to run goplay right after installing it