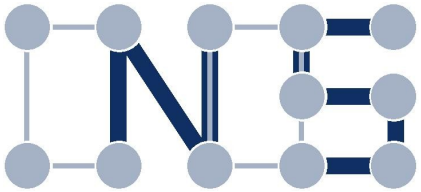
JACOBS
UNIVERSITY



Crawling Bug Tracker for Semantic Bug Search

Ha Manh Tran, Georgi Chulkov and Jürgen Schönwälder
Computer Science, Jacobs University Bremen, Germany

19th IFIP/IEEE International Workshop on
Distributed Systems: Operations and Management
Samos Island, Greece , 25-26 September 2007

Overview

- Purpose
  - Study of semantic search on bug reports from bug tracking systems

- Focus
  - Bug tracking systems
  - Bug crawling
  - Unified data model
  - Semi-structured bug search

Bug Tracking System (BTS)

- BTS "focal" features (as of October 2007)

Tracker	License	Access	Updates	Schema	Dep.†	Search
Bugzilla	MPL	HTML,XML-RPC*	SMTP, RSS*	textual	optional	filter,keywords
Mantis	GPL	HTML, SOAP*	SMTP, RSS	graphical	yes	filter
Trac	BSD	HTML	SMTP, RSS*	graphical	no	filter,keywords
Debian BTS	GPL	HTML, SMTP	SMTP	unknown	optional	filter
phpBugTracker	GPL	HTML	SMTP	textual	yes	filter,keywords
Flyspray	LGPL	HTML	SMTP,RSS,XMPP	unknown	yes	filter,keywords

  - Licenses affect BTS popularity
  - Web service interfaces facilitate fast retrieval
  - Schema exposes the structure of bug reports
  - Bug dependency and search options keep track of related bug reports

Popular BTS

- BTS popularity (as of October 2007)

Site	System	Version	Bugs	Activity	Custom	RPC	RSS	Dep.†
bugs.debian.org	Debian BTS	N/A	349346	1036	N/A	no	no	no
bugs.kde.org	Bugzilla	unknown	9655+	24+	light	no	no	no
bugs.eclipse.org	Bugzilla	unknown	204600	746	heavy	yes	yes	yes
bugs.gentoo.org	Bugzilla	unknown	183365	538	none	no	yes	yes
bugzilla.mozilla.org	Bugzilla	3.0.1+	173885	721	none	yes	yes	yes
bugzilla.redhat.com	Bugzilla	2.18-rh	177724	unknown	light	yes	yes	yes
qa.netbeans.org	Bugzilla	unknown	116639+	unknown	heavy	no	no	yes
bugs.digium.com	Mantis	unknown	10765	63	none	no	yes	yes
bugs.scribus.net	Mantis	1.0.7	6142	24	none	no	yes	yes
bugtrack.alsa-project.org	Mantis	1.0.6	3430	22	none	no	no	yes
dev.rubyonrails.org	Trac	0.10.5dev	11493+	unknown	none	no	yes	no
trac.edgewall.org	Trac	unknown	5948	unknown	none	no	yes	no
bugs.icu-project.org	Trac	0.10.4	5845	unknown	none	no	yes	no

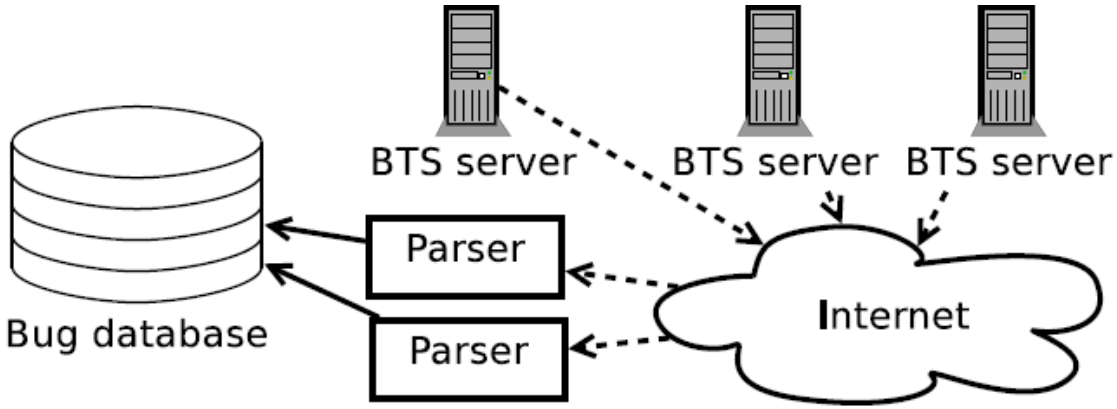
  - The number of bug reports and the activity of a site imply the popularity of the BTS system

Crawling BTS

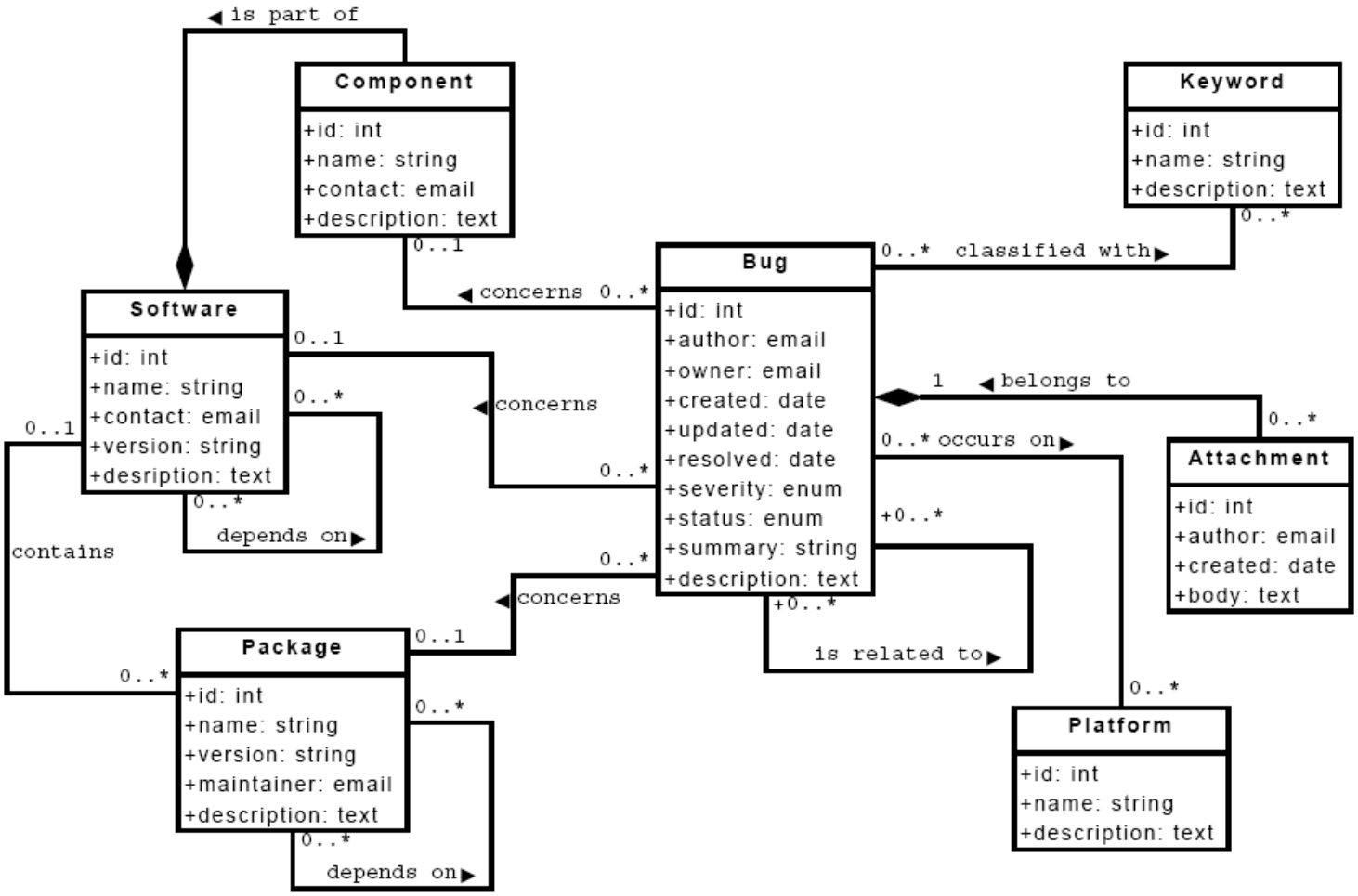
- Bugzilla BTS
  - Retrieve a large number of bug reports
  - Support *XML-RPC* APIs
  - Ignore any attachment of bug reports
- Debian BTS
  - Access the raw bug data directly
  - Support *rsync* utility to copy bug reports
  - Separate active and closed bug reports

Crawling BTS

- Buglook tool
  - Access bug reports via a web interface: download and parse HTML pages
  - Most BTSs support the web interface
  - Web pages contain more complete data
  - Customized web pages can cause difficulties



Unified Data Model



Unified Data Model

- Integrating bug reports from different BTSs

Unified model	Bugzilla	Trac	Mantis	Debian
critical	blocker, critical	blocker, critical	block, crash	critical, grave, serious
normal	major	major	major	important, normal
minor	minor, trivial	minor, trivial	minor, tweak, text, trivial	minor
feature	enhancement	-	feature	wishlist

- Including necessary details for search purpose
  - Administrative meta-data
  - Bug description and attachment
- Focusing on relationships
  - Keyword/symptom classification, package/software dependency, bug relation.

Multiple Vector Representation

- Field-value vector
  - Pre-defined fields of status information
  - Fields of classification and diagnosis information
  - Meta-data search using keyword extraction and evaluation
- Semantic vector
  - Textual description of a problem
  - Full-text search using algebraic computation
- Good performance on the CISI and MED bibliographic datasets

Experiment Metric

- Matching rate

$$r = \frac{|S_x \cap S_y|}{min(N_x, N_y)}$$

  - $|S_x|$ the length of the resulting set of bugs obtained by algorithm X
  - N_x the total number of bugs obtained by algorithm X

- Three combinations of search algorithms
  - Full-text search and meta-data search (ft-md)
  - Full-text search and keyword search (ft-bl)
  - Meta-data search and keyword search (md-bl)

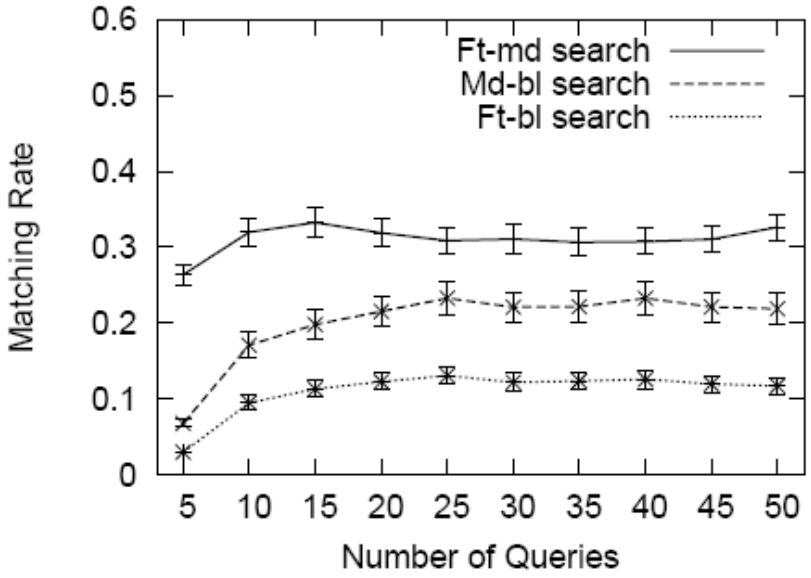
Experiment Setup

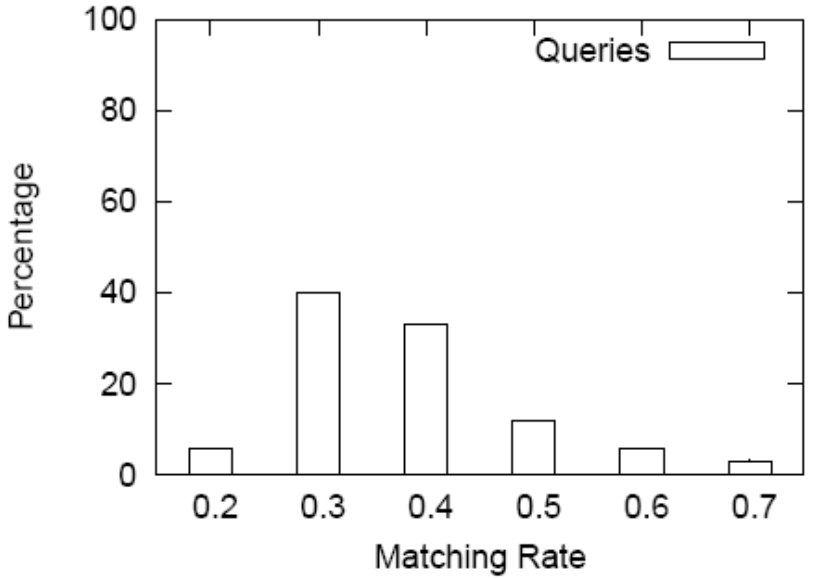
- Figures
  - Data set contains 11.077 bugs
  - Query set contains 50 queries
  - Number of bugs obtained is 100 bugs (N_x)

- Implementations
  - Porter stemming algorithm
  - Single-vector Lanczos algorithm
  - Testing on x86 64 GNU/Linux machine with two dual-core AMD Opteron(tm) processors running at 2 GHz with 4 GB RAM

Search Performance

- Performance by average matching rate

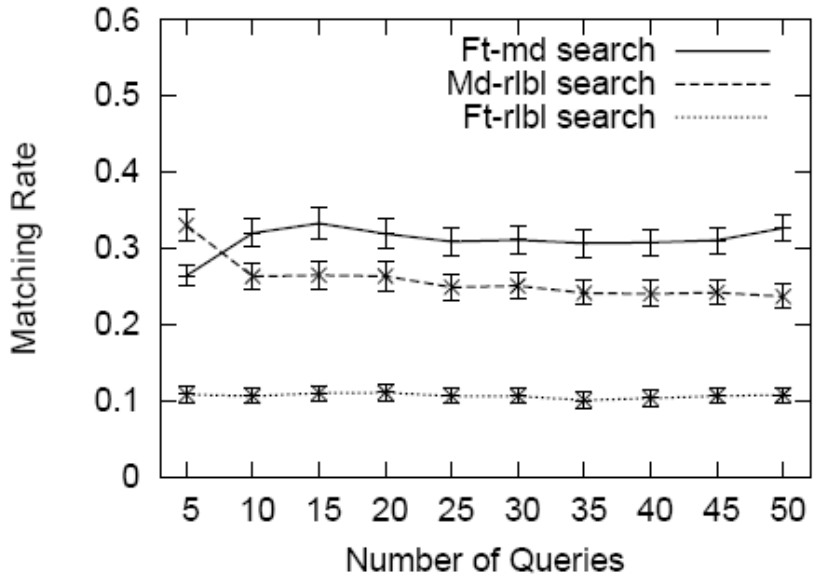


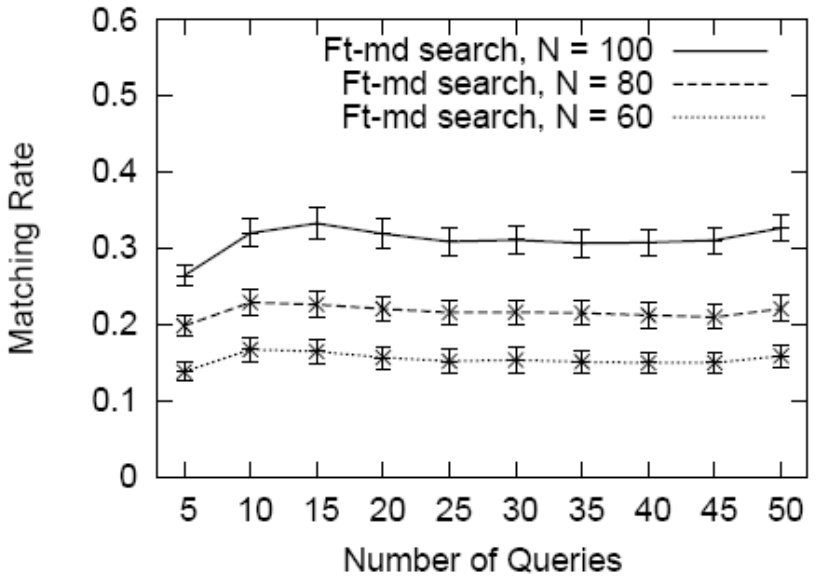


- Ft-md provides more identical bugs than md-bl and ft-bl
- The average matching rate is low (0.3)

Search Performance

- Performance of different N_x values





- Ft-md fits well for semi-structured bug data
- The data set is wide and diverse in scope and the number of duplicated bugs is small

Conclusions

- Bug reports from BTSs are invaluable
  - Problem resolution, problem search
- We crawled bug reports from popular BTSs and created a unified data model
  - Bug dataset
- We facilitated semantic search on bug reports using MVR
- Future work focuses on a complete online semantic search system

Thank you for your attention.

Questions?